

UNIVERSITÉ DE SHERBROOKE
Faculté de génie
Département de génie électrique et de génie informatique

RECONNAISSANCE DE LOCUTEURS POUR ROBOT MOBILE

Mémoire de maîtrise
Spécialité : génie électrique

François GRONDIN

Jury : François MICHAUD, ing. Ph.D. (directeur)
Roch LEFEBVRE, ing. Ph.D.
Pierre DUMOUCHEL, ing. Ph.D.

À mon père, Robert.

RÉSUMÉ

L’audition artificielle est de plus en plus utilisée en robotique mobile pour améliorer l’interaction humain-robot. La reconnaissance de la parole occupe présentement une place importante tandis qu’un intérêt particulier se développe pour la reconnaissance de locuteurs. Le système ManyEars permet actuellement à un robot mobile de localiser, suivre et séparer plusieurs sources sonores. Ce système utilise un ensemble de huit microphones qui sont disposés en cube. Ce mémoire porte sur la conception et l’évaluation d’un système de reconnaissance de locuteurs, baptisé WISS (*Who IS Speaking*), couplé au système ManyEars. Le système de reconnaissance de locuteurs conçu est robuste au bruit ambiant et au changement d’environnement. Une technique de combinaison de modèle parallèle (*parallel model combination* (PMC)) et des masques sont utilisés pour améliorer le taux d’identification dans un milieu bruité. Un indice de confiance est également introduit pour pondérer les identifications obtenues. La simplicité du système proposé fait en sorte qu’il est possible d’exécuter en temps réel l’algorithme sur un processeur généraliste (*General Purpose Processor* (GPP)).

Les performances du système sont établies à l’aide de plusieurs scénarios. Dans un premier lieu, des enregistrements sont diffusés dans des haut-parleurs pour un ensemble de vingt locuteurs. Le système est ainsi caractérisé en fonction des positions angulaires et radiales des sources sonores. Le taux de reconnaissance est affecté par la qualité du signal (i.e. diminution du rapport signal sur bruit (*Signal-to-Noise Ratio* (SNR))) : il passe de 95.6% à 84.3% en moyenne lorsque le SNR passe d’environ 16 dB à 2 dB lorsque le locuteur se situe à 1.5 mètres des microphones. Par la suite, un scénario dit statique est vérifié à l’aide de quatre locuteurs qui récitent chacun leur tour des phrases à un volume de voix naturel. Finalement, un scénario dynamique dans lequel un groupe de quatre locuteurs ont une conversation naturelle avec des chevauchements entre les segments de paroles est étudié. Le taux de reconnaissance varie entre 74.2% et 100.0% (avec une moyenne de 90.6%) avec le scénario statique, et entre 42.6% et 100.0% avec le scénario dynamique (avec des moyennes de 58.3%, 72.8% et 81.4% pour des segments de 1, 2 et 3 secondes respectivement). Des solutions sont identifiées afin d’améliorer les performances lors de travaux futurs.

Au meilleur de notre connaissance, il n’existe aucun système qui effectue une reconnaissance de locuteurs dans un environnement contaminé simultanément par des bruits convolutif et additif. De plus, l’utilisation de masques pour estimer ces bruits est un nouveau concept. Ces masques sont d’ailleurs généralement employés pour la reconnaissance de la parole et leur utilisation dans un contexte de reconnaissance de locuteur est une première. De plus, une caractérisation complète du système qui inclue les SNRs est proposée en fonction de la position du locuteur, ce qui est rarement disponible dans la littérature en audition artificielle pour les robots.

Mots-clés : Audition artificielle, robot mobile, reconnaissance de locuteur, environnement dynamique, bruits additif et convolutif

REMERCIEMENTS

Je tiens d'abord à remercier mon directeur François Michaud pour son soutien et sa disponibilité tout au long de ma maîtrise. La confiance qu'il m'a manifestée a été un élément-clé dans la réussite de ce projet. Tout au long de mon parcours, il a su me transmettre sa passion pour le domaine, ce qui a consolidé mon intérêt pour la recherche. François Michaud est titulaire de la Chaire de recherche en robotique mobile et systèmes intelligents autonomes à l'Université de Sherbrooke. Il est également le directeur de l'Institut interdisciplinaire d'innovation technologique (3IT) de l'Université de Sherbrooke.

Durant ce projet, j'ai été soutenu financièrement par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) et le Fonds québécois de la recherche sur la nature et les technologies (FQRNT). Je tiens à remercier ces organismes pour leur contribution significative.

J'aimerais remercier Jean-Marc Valin pour ses réponses rapides et précises à mes questions en lien avec le système ManyEars. Je suis également reconnaissant auprès du professeur Roch Lefebvre de l'Université de Sherbrooke qui a généreusement mis à ma disposition une base de données de locuteurs. Cette dernière s'est avérée être un outil essentiel durant les expériences en laboratoire.

De plus, je tiens à remercier Dominic Létourneau pour son aide avec le matériel utilisé pour effectuer l'acquisition des signaux audio. Je tiens également à souligner les conseils judicieux de François Ferland et Mathieu Labbé concernant la programmation du système en langage C. J'aimerais aussi mentionner la contribution spéciale de Vincent Rousseau pour avoir intégré la librairie ManyEars dans l'environnement ROS. Finalement, j'aimerais remercier tous les membres du laboratoire IntRoLab qui ont participé aux expériences audio en laboratoire.

TABLE DES MATIÈRES

1	INTRODUCTION	1
2	RECONNAISSANCE DE LOCUTEURS EN ROBOTIQUE	3
3	MANYEARS, UN SYSTÈME DE LOCALISATION, SUIVI ET SÉPARATION	9
3.1	Localisation de sources sonores avec ManyEars	12
3.1.1	Analyse et FFT	13
3.1.2	Pondération et corrélation croisée	14
3.1.3	Sphère unitaire	17
3.1.4	Recherche	20
3.2	Suivi de sources sonores avec ManyEars	22
3.2.1	Prédiction	23
3.2.2	Probabilité instantanée	24
3.2.3	Probabilité pour des sources multiples	24
3.2.4	Mise à jour de la pondération des particules	29
3.2.5	Ajout d'une nouvelle source	29
3.2.6	Suppression d'une source existante	30
3.2.7	Estimation de la position des sources	30
3.2.8	Échantillonnage des particules	30
3.3	Séparation de sources sonores avec ManyEars	31
3.4	Post-filtrage sur ManyEars	34
3.4.1	Estimation du bruit	35
3.4.2	Gain lorsqu'une source est active	36
3.4.3	Gain selon l'activité vocale	38
4	WISS, UN SYSTÈME DE RECONNAISSANCE DE LOCUTEURS	41
4.1	Entraînement de WISS	45
4.1.1	Analyse et FFT	45
4.1.2	Caractéristiques vocales	46
4.1.3	Modèle	52
4.2	Reconnaissance de locuteurs avec WISS	53
4.2.1	Caractéristiques vocales	54
4.2.2	Mise à jour des modèles	66
4.2.3	Pointage et prise de décision	67
4.2.4	Indice de confiance	68
4.3	Contributions	69
5	ANALYSE DE PERFORMANCES	71
5.1	Caractérisation du système	73
5.1.1	Performances selon la position angulaire	74

5.1.2	Performances selon la position radiale	79
5.1.3	Performances sur l'ensemble des positions	83
5.2	Interaction statique	84
5.3	Interaction dynamique	86
5.4	Considérations temps réel	88
5.5	Discussion	90
6	CONCLUSION	93
A	INTERACTION DYNAMIQUE	95
A.1	Courts segments	96
A.2	Moyens segments	102
A.3	Longs segments	108
	LISTE DES RÉFÉRENCES	115

LISTE DES FIGURES

2.1	Modèle de la parole	4
3.1	Schéma-bloc du système ManyEars	10
3.2	Aperçu de la localisation avec ManyEars	12
3.3	Fenêtre d'analyse et de synthèse	14
3.4	Aperçu du mécanisme de suivi de ManyEars	22
3.5	Aperçu de la séparation de sources sonores avec ManyEars	31
3.6	Aperçu du post-filtrage de sources sonores avec ManyEars	35
3.7	Fonction transcendante du gain pour une source active	37
4.1	Modélisation des bruits convolutif et additif	42
4.2	Schéma-bloc des modules pour l'entraînement de WISS	45
4.3	Exemple du logarithme du spectre avec emphase pour l'entraînement	47
4.4	Exemple de l'activité vocale pour l'entraînement	49
4.5	Banc de filtres	50
4.6	Exemple des caractéristiques vocales pour l'entraînement	51
4.7	Exemple des caractéristiques vocales actives pour l'entraînement	51
4.8	Exemple des caractéristiques vocales normalisées	52
4.9	Exemple de modélisation des caractéristiques vocales par quantification vectorielle	52
4.10	Schéma-bloc des modules pour la reconnaissance de locuteurs avec WISS	53
4.11	Exemple du logarithme du spectre des signaux pour l'évaluation	55
4.12	Exemple de caractéristiques vocales pour l'évaluation	56
4.13	Exemple de masques instantanés	57
4.14	Exemple de l'estimation du bruit additif	58
4.15	Exemple de masques en fréquence	58
4.16	Exemple de l'estimation du bruit convolutif	63
4.17	Exemple de masques temporels	63
4.18	Exemple de caractéristiques vocales normalisées dans le domaine spectral	64
4.19	Combinaison des masques instantanés, temporels et en fréquence pour obtenir les masques globaux	65
4.20	Exemple d'un masque global	65
4.21	Modification dynamique des centroïdes	67
4.22	Courbe de fonction sigmoïde pour l'indice de confiance	68
5.1	Microphones et système d'acquisition	71
5.2	Positions angulaires	75
5.3	Performances avec les indices non-pondérés pour les huit positions angulaires avec le formateur de faisceaux sans masques	77
5.4	Performances avec les indices pondérés pour les huit positions angulaires avec le formateur de faisceaux sans masques	77

5.5	Performances avec les indices non-pondérés pour les huit positions angulaires avec le formateur de faisceaux avec masques	78
5.6	Performances avec les indices pondérés pour les positions angulaires avec le formateur de faisceaux avec masques	78
5.7	Positions radiales	79
5.8	Performances avec les indices non-pondérés pour les positions radiales avec le formateur de faisceaux sans masques	81
5.9	Performances avec les indices pondérés pour les positions radiales avec le formateur de faisceaux sans masques	81
5.10	Performances avec les indices non-pondérés pour les positions radiales avec le formateur de faisceaux avec masques	82
5.11	Performances avec les indices pondérés pour les positions radiales avec le formateur de faisceaux avec masques	82
5.12	Moyennes des taux de bonnes identifications selon la méthode sélectionnée	83
5.13	Positions des locuteurs pour le scénario statique	85
5.14	Performances avec les indices pondérés pour chaque locuteur avec masques dans un contexte d'interaction statique	85
5.15	Positions des locuteurs pour le scénario dynamique	86
5.16	Performances avec les indices pondérés pour chaque locuteur avec masques dans un contexte d'interaction dynamique	88
5.17	Fils d'exécution pour un scénario en temps réel	89
6.1	Positions des microphones sur le robot Johnny-0	94
A.1	Positions des locuteurs pour le scénario dynamique	95
A.2	Identifications pour des segments de courte durée (0-60 secs)	96
A.3	Identifications pour des segments de courte durée (60-120 secs)	96
A.4	Identifications pour des segments de courte durée (120-180 secs)	97
A.5	Identifications pour des segments de courte durée (180-240 secs)	97
A.6	Identifications pour des segments de courte durée (240-300 secs)	98
A.7	Identifications pour des segments de courte durée (300-360 secs)	98
A.8	Identifications pour des segments de courte durée (360-420 secs)	99
A.9	Identifications pour des segments de courte durée (420-480 secs)	99
A.10	Identifications pour des segments de courte durée (480-540 secs)	100
A.11	Identifications pour des segments de courte durée (540-600 secs)	100
A.12	Identifications pour des segments de courte durée (600-660 secs)	101
A.13	Identifications pour des segments de courte durée (660-720 secs)	101
A.14	Identifications pour des segments de durée moyenne (0-60 secs)	102
A.15	Identifications pour des segments de durée moyenne (60-120 secs)	102
A.16	Identifications pour des segments de durée moyenne (120-180 secs)	103
A.17	Identifications pour des segments de durée moyenne (180-240 secs)	103
A.18	Identifications pour des segments de durée moyenne (240-300 secs)	104
A.19	Identifications pour des segments de durée moyenne (300-360 secs)	104
A.20	Identifications pour des segments de durée moyenne (360-420 secs)	105
A.21	Identifications pour des segments de durée moyenne (420-480 secs)	105

A.22	Identifications pour des segments de durée moyenne (480-540 secs)	106
A.23	Identifications pour des segments de durée moyenne (540-600 secs)	106
A.24	Identifications pour des segments de durée moyenne (600-660 secs)	107
A.25	Identifications pour des segments de durée moyenne (660-720 secs)	107
A.26	Identifications pour des segments de longue durée (0-60 secs)	108
A.27	Identifications pour des segments de longue durée (60-120 secs)	108
A.28	Identifications pour des segments de longue durée (120-180 secs)	109
A.29	Identifications pour des segments de longue durée (180-240 secs)	109
A.30	Identifications pour des segments de longue durée (240-300 secs)	110
A.31	Identifications pour des segments de longue durée (300-360 secs)	110
A.32	Identifications pour des segments de longue durée (360-420 secs)	111
A.33	Identifications pour des segments de longue durée (420-480 secs)	111
A.34	Identifications pour des segments de longue durée (480-540 secs)	112
A.35	Identifications pour des segments de longue durée (540-600 secs)	112
A.36	Identifications pour des segments de longue durée (600-660 secs)	113
A.37	Identifications pour des segments de longue durée (660-720 secs)	113

LISTE DES TABLEAUX

3.1	Points initiaux pour la sphère icosaédrale	18
3.2	Triangles initiaux pour la sphère icosaédrale	19
3.3	Paramètres des particules	24
3.4	Exemple pour l'ensemble G^l	25
3.5	Fenêtres utilisées pour la détection de l'activité vocale	39
4.1	Banc de filtres	49
5.1	Positions des microphones (en mètres)	72
5.2	SNR selon la position angulaire	76
5.3	SNR selon la position radiale	80
5.4	Durée des segments et taux de bonnes identifications	87

LISTE DES ACRONYMES

Acronyme	Définition
ABMMIC	Séparation avec un critère minimum d'information mutuelle (<i>Adaptive Beamforming with a Minimum Mutual Information Criterion</i>)
CMN	Normalisation des moyennes cepstrales (<i>Cepstral Mean Normalization</i>)
DSP	Processeur de signal numérique (<i>Digital Signal Processor</i>)
FFT	Transformée de Fourier rapide (<i>Fast Fourier Transform</i>)
GMM	Modèle de mélange gaussien (<i>Gaussian Mixture Model</i>)
GPP	Processeur généraliste (<i>General Purpose Processor</i>)
GSC	Éliminateur de lobes latéraux (<i>Generalized Sidelobe Canceller</i>)
GSS	Séparation géométrique des sources (<i>Geometric Source Separation</i>)
HMM	Modèle de Markov caché (<i>Hidden Markov Model</i>)
ICA	Analyse en composantes indépendantes (<i>Independant Component Analysis</i>)
IFFT	Transformée de Fourier rapide inverse (<i>Inverse Fast Fourier Transform</i>)
LPC	Coefficient à prédiction linéaire (<i>Linear Predictive Coefficient</i>)
MCRA	Moyenne récursive obtenue à l'aide des minima (<i>Minima-Controlled Recursive Average</i>)
MFCC	Coefficient cepstral réparti sur l'échelle Mel (<i>Mel-Frequency Cepstral Coefficient</i>)
PMC	Combinaison de modèle parallèle (<i>Parallel Model Combination</i>)
SNR	Rapport signal sur bruit (<i>Signal-to-Noise Ratio</i>)
SSE	Extension pour flux SIMD (<i>Streaming SIMD Extension</i>)
VQ	Quantification vectorielle (<i>Vector Quantization</i>)

Variable	Définition
$a_{m,s}^l[k]$	Délai d'arrivée dans le domaine fréquentiel pour la séparation pour le microphone m , la source s , l'indice fréquentiel k et la trame l
$a_s^l(f)$	Paramètre d'amortissement associé à une particule f au sein d'un filtre s pour la trame l
$\mathbf{A}^l[k]$	Matrice d'arrivée des signaux pour la séparation à l'indice fréquentiel k et la trame l
$b[k]$	Fenêtre utilisée pour l'estimation du bruit stationnaire à l'indice fréquentiel k
$b[t]$	Bruit additif total à l'indice temporel t
$b_{back,m}[t]$	Bruit ambiant capté par le microphone m à l'indice temporel t
$(b_{center})_\Lambda$	Fréquence centrale du filtre Λ
$(b_{max})_\Lambda$	Fréquence maximale du filtre Λ
$(b_{min})_\Lambda$	Fréquence minimale du filtre Λ
$\mathbf{b}_{noisy}[t]$	Ensemble des bruits additifs des microphones à l'indice temporel t
$b_s^l(f)$	Paramètre d'excitation associé à une particule f au sein d'un filtre s pour la trame l
$b_\Lambda[k]$	Échantillon à l'indice fréquentiel k du filtre Λ
$ B(j\omega) ^2$	Densité spectrale de puissance du bruit additif à l'indice fréquentiel ω
$\hat{\mathbf{B}}_r$	Estimation du bruit additif pour le segment r
$\hat{B}_r[\Lambda]$	Estimation du bruit additif au filtre Λ pour le segment r
c	Indice de la réalisation pour des sources multiples
c_{air}	Vitesse du son dans l'air
c_{frame}	Indice de l'échantillon de la fenêtre entière
c_{global}	Indice de l'échantillon de la fenêtre large
c_{local}	Indice de l'échantillon de la fenêtre étroite
$(c_{train}^{speech})_u^l[\Lambda]$	Caractéristiques d'entraînement avec une activité vocale non nulle du locuteur u à la trame l au filtre Λ
$(c_{test})_r^l[\Lambda]$	Caractéristiques de test du segment r à la trame l au filtre Λ
$(c_{test}^{ac})_r^l[\Lambda]$	Caractéristiques de test normalisées du segment r à la trame l au filtre Λ
$(c_{test}^{speech})_r^l[\Lambda]$	Caractéristiques de test avec une activité vocale non nulle du segment r à la trame l au filtre Λ
$(\mathbf{c}_{test}^{speech})_r^l$	Caractéristiques de test avec une activité vocale non nulle du segment r à la trame l pour tous les filtres

Variable	Définition
$(c_{train})_u^l[\Lambda]$	Caractéristiques d'entraînement du locuteur u à la trame l au filtre Λ
$(c_{train}^{cmn})_u^l[\Lambda]$	Caractéristiques d'entraînement normalisées du locuteur u à la trame l au filtre Λ
$conf_r$	Indice de confiance du segment r
C	Nombre de réalisations pour des sources multiples
d	Indice d'un point sur la sphère unitaire
$(d_{max})_q^l$	Indice du point sur la sphère possédant l'énergie maximale pour la source potentielle q à la trame l
$delay_{m_1,m_2}(d)$	Délai d'arrivée du signal entre les microphones m_1 et m_2 lorsque la source se situe au point d sur la sphère unitaire
D_{level}	Nombre de points sur la sphère unitaire au niveau $level$
$\mathbf{E}^l[k]$	Matrice de corrélation pour les signaux des sources séparées sans autocorrélation pour la trame l
E_{disp}	Dispersion des particules
$(E_{pot})_q^l$	Énergie de la source potentielle q pour la trame l
$E_q^l(d)$	Énergie du formateur de faisceaux de la source potentielle q au point d sur la sphère unitaire pour la trame l
E_T	Seuil d'énergie des sources potentielles
f	Indice d'une particule au sein d'un filtre particulière
f_{sample}	Nouvel indice d'une particule suite à l'échantillonnage
F	Nombre de particules au sein d'un filtre particulière
F_s	Taux d'échantillonnage
F_x	Réalisation d'une variable aléatoire avec une distribution normale
\mathbf{g}_c^l	Réalisation c des relations entre toutes les sources potentielles et des sources suivies pour une trame l
$g_c^l(q)$	Réalisation c de la relation entre une source potentielle q et des sources suivies pour une trame l
G	Volume du locuteur
G^l	Ensemble des réalisations pour l'attribution des hypothèses pour toutes les sources potentielles
G_{min}	Gain minimal pour le post-filtrage
$G_s^l[k]$	Gain pour le post-filtrage à l'indice fréquentiel k de la source post-filtrée s au sein de la trame l
$(G_t)_s^l[k]$	Gain représenté par une fonction transcendante pour le post-filtrage à l'indice fréquentiel k de la source s à la trame l

Variable	Définition
$(G_{H_1})_s^l[k]$	Gain pour le post-filtrage lorsqu'une source est active à l'indice fréquentiel k de la source s à la trame l
$h[t]$	Bruit convolutif total à l'indice temporel t
$h_{frame}[c_{frame}]$	Fenêtre entière de filtrage Hann pour le post-filtrage
$h_{global}[c_{global}]$	Fenêtre large de filtrage Hann pour le post-filtrage
$h_{local}[c_{local}]$	Fenêtre étroite de filtrage Hann pour le post-filtrage
$h_{mic,m}[t]$	Réponse impulsionnelle pour le microphone m
$\mathbf{h}_{noisy}[t]$	Ensemble des bruits convolutifs des microphones à l'indice temporel t
$h_{room,m}[t]$	Réponse impulsionnelle de la pièce entre la source et le microphone m à l'indice temporel t
$ H(j\omega) ^2$	Densité spectrale de puissance du bruit convolutif à l'indice fréquentiel ω
$H_{emph}[k]$	Puissance du spectre du filtre passe-haut à l'indice fréquentiel k
\mathbf{H}_r	Bruit convolutif pour le segment r
$\hat{\mathbf{H}}_r$	Estimation du bruit convolutif pour le segment r
$\hat{\mathbf{H}}_r^{u,v}$	Estimation du bruit convolutif pour le segment r , le locuteur u et le centroïde v
$(id_{exp})_r$	Identité obtenue pour le segment r
$(id_{theo})_r$	Identité réelle du segment r
$J_1(\mathbf{W}^l[k])$	Contrainte statistique pour la séparation à l'indice fréquentiel k et la trame l
$J_2(\mathbf{W}^l[k])$	Contrainte géométrique pour la séparation à l'indice fréquentiel k et la trame l
k	Indice fréquentiel d'un échantillon au sein d'une trame
l	Indice de la trame
$level$	Indice du niveau d'itération sur la sphère unitaire
$(L_{noise})_s$	Nombre de trames bruitées pour la source s
$(L_{speech})_s$	Nombre de trames de parole pour la source s
L_{conf}	Nombre de trames pour confirmer l'existence d'une source
L_{delete}	Nombre de trames pour supprimer une source
L_{total}	Nombre de trames pour le signal de chaque microphone
L_{mcra}	Nombre de trames utilisées pour évaluer le bruit stationnaire
$(L_{test})_r$	Nombre de trames pour le segment de test r

Variable	Définition
$(L_{test}^{speech})_r$	Nombre de trames dont le masque binaire en fréquence est non nul pour le segment de test r
$(L_{train}^{all})_u$	Nombre total de trames utilisées pour l'entraînement du locuteur u
$(L_{train}^{speech})_u$	Nombre de trames de parole pour le locuteur u
m	Indice d'un microphone
$(m_{all})_r^l[\Lambda]$	Masque binaire global au filtre Λ pour le segment r à la trame l
$(m_{freq})_r^l$	Masque binaire en fréquence pour le segment r à la trame l
$(m_{inst})_r^l[\Lambda]$	Masque binaire instantané au filtre Λ pour le segment r à la trame l
$(m_{time})_r[\Lambda]$	Masque binaire temporel pour le segment r
$(match_{bad}^{noWeight})_r$	Nombre de mauvaises identifications non-pondéré
$(match_{bad}^{weight})_r$	Nombre de mauvaises identifications pondéré
$(match_{good}^{noWeight})_r$	Nombre de bonnes identifications non-pondéré
$(match_{good}^{weight})_r$	Nombre de bonnes identifications pondéré
M	Nombre total de microphones
n	Indice temporel d'un échantillon au sein d'une trame
N	Nombre d'échantillons par trame
o	Facteur de chevauchement entre chaque trame
p_{acc}	Probabilité pour qu'une particule soit en accélération
p_{cst}	Probabilité pour qu'une particule se déplace à vitesse constante
p_{stp}	Probabilité pour qu'une particule soit en arrêt
$(p_{speech})_s^l[k]$	Probabilité que la source s soit active à l'indice fréquentiel k et la trame l
$p((\mathbf{z}_{part})_s^l(f) O^l)$	La probabilité que la particule f du filtre s soit à la position $(\mathbf{z}_{part})_s^l(f)$ suite à l'observation des sources potentielles O^l
$p(O^l \mathbf{g}_c^l)$	Probabilité que l'ensemble des sources potentielles observées correspondent à la réalisation c de la relation de toutes les sources potentielles et des sources suivies pour une trame l
$p(O_q^l g_c^l(q))$	Probabilité que la source potentielle q observée correspondent à la réalisation c de la relation entre une source potentielle q et des sources suivies pour une trame l
$p(O_q^l (\mathbf{z}_{part})_s^l(f))$	La probabilité d'observer une source potentielle O_q^l à la position d'une particule $(\mathbf{z}_{part})_s^l(f)$
P_{false}	Probabilité <i>a priori</i> qu'une source potentielle soit une fausse détection

Variable	Définition
$(P_{frame})_s^l[k]$	Probabilité d'activité vocale sur une fenêtre entière à l'indice fréquentiel k de la source s à la trame l
$(P_{frame})_u^l[k]$	Probabilité d'activité vocale sur une fenêtre entière à l'indice fréquentiel k du locuteur u à la trame l
$P(\mathbf{g}_c^l)$	Probabilité d'obtenir la réalisation c de la relation entre toutes les sources potentielles et des sources suivies pour une trame l
$P(\mathbf{g}_c^l O^l)$	Probabilité d'obtenir la réalisation de \mathbf{g}_c^l si l'observation O^l est connue
$P(g_c^l(q))$	Probabilité d'obtenir la réalisation c de la relation entre une source potentielle q et des sources suivies pour une trame l
$(P_{global})_s^l[k]$	Probabilité d'activité vocale sur une fenêtre large à l'indice fréquentiel k de la source s à la trame l
$(P_{global})_u^l[k]$	Probabilité d'activité vocale sur une fenêtre large à l'indice fréquentiel k du locuteur u à la trame l
$(P_{local})_s^l[k]$	Probabilité d'activité vocale sur une fenêtre étroite à l'indice fréquentiel k de la source s à la trame l
$(P_{local})_u^l[k]$	Probabilité d'activité vocale sur une fenêtre étroite à l'indice fréquentiel k du locuteur u à la trame l
P_b	Probabilité minimum qu'une source suivie s soit active à la trame $l - 1$ selon les observations de la trame $l - 1$ seulement
P_m	Mise à l'échelle de la probabilité qu'une source suivie s soit active à la trame $l - 1$ selon les observations de la trame $l - 1$ seulement
P_{new}	Probabilité <i>a priori</i> qu'une nouvelle source apparaisse
P_o	Probabilité <i>a priori</i> qu'une source suivie ne soit pas observée malgré le fait qu'elle existe
P_q	Probabilité que la source potentielle q existe réellement
$(P_s^l)(q)$	Probabilité normalisée que le filtre s soit associé à la source potentielle q à la trame l
$(P_s^l)'(q)$	Probabilité non normalisée que le filtre s soit associé à la source potentielle q à la trame l
P_s^l	Probabilité qu'une source s déjà suivie à la trame $l - 1$ soit associée à une source potentielle
P_s^{l-1}	Probabilité que la source suivie s à la trame l soit déjà suivie depuis la trame précédente
P_{tot}	Somme des probabilités pour toutes les hypothèses
$(P_{vad})_u^l$	État de l'activité vocale pour le locuteur u à la trame l

Variable	Définition
$P(A_s^l \neg A_s^{l-1})$	Probabilité qu'une source suivie s active à la trame $l - 1$ demeure inactive à la trame l
$P(A_s^l A_s^{l-1})$	Probabilité qu'une source suivie s active à la trame $l - 1$ demeure active à la trame l
$P(A_s^l \mathbf{O}^{l-1})$	Probabilité qu'une source suivie s soit active à la trame l
$P(A_s^{l-1} \mathbf{O}^{l-1})$	Probabilité qu'une source suivie s soit active à la trame $l - 1$ selon les observations de la trame $l - 1$ seulement
$P(E_s^l \mathbf{O}^{l-1})$	Probabilité qu'une source suivie s existe à la trame l
$P(Ob_s^l \mathbf{O}^{l-1})$	Probabilité que la source suivie s soit observée à la trame l
$(P_{H_0}^l)(q)$	Probabilité normalisée que l'hypothèse H_0 soit valide pour la source potentielle q à la trame l
$(P_{H_0}^l)'(q)$	Probabilité non normalisée que l'hypothèse H_0 soit valide pour la source potentielle q à la trame l
$(P_{H_2}^l)(q)$	Probabilité normalisée que l'hypothèse H_2 soit valide pour la source potentielle q à la trame l
$(P_{H_2}^l)'(q)$	Probabilité non normalisée que l'hypothèse H_2 soit valide pour la source potentielle q à la trame l
q	Indice d'une source potentielle
$(q_{speech})_s^l[k]$	Probabilité <i>a priori</i> que la source s ne soit pas active à l'indice fréquentiel k et la trame l
Q	Nombre de sources potentielles
r	Indice du segment de test
$rate_{noWeight}$	Taux d'identification non-pondéré
$rate_{weight}$	Taux d'identification pondéré
$(ratio_{inst})_r^l[\Lambda]$	Rapport entre la puissance du spectre post-filtré et le spectre séparé pour le segment r à la trame l
$(ratio_{time})_r[\Lambda]$	Différence entre la moyenne des segments de parole et le bruit additif pour le segment r
$R_{m_1, m_2}^l(\tau)$	Corrélation croisée entre les signaux des microphones m_1 et m_2 pour un délai τ à la trame l
$\mathbf{R}_{mm}^l[k]$	Matrice de corrélation pour les signaux des microphones
$\mathbf{R}_{ss}^l[k]$	Matrice de corrélation pour les signaux des sources séparées
s	Indice d'un filtre particulier
S	Nombre de filtres particuliers
s_{new}	Indice d'un nouveau filtre
$score_u^r$	Pointage du segment r par rapport au modèle du locuteur u

Variable	Définition
t	Indice temporel d'un échantillon
T_{conf}	Seuil pour confirmer l'existence d'une source
T_{delete}	Seuil pour supprimer un filtre particulière
T_{disp}	Seuil pour échantillonner les particules d'un filtre
T_{freq}	Seuil pour le masque binaire en fréquence
T_{inst}	Seuil pour le masque binaire instantané
$T_{minFeatures}$	Seuil pour la durée minimum en termes des caractéristiques vocales valides
T_{new}	Seuil pour créer un nouveau filtre particulière
T_{vad}	Seuil d'état de l'activité vocale
u	Indice du locuteur
U	Nombre de locuteurs
v	Indice du centroïde
v_q	Rapport entre l'énergie d'une source potentielle q et le seuil d'énergie
$v_s^l[k]$	Fonction simplifiée pour le calcul du gain de post-filtrage à l'indice fréquentiel k de la source s à la trame l
V	Nombre de centroïdes
$w[n]$	Échantillon des fenêtres d'analyse et de synthèse à l'indice temporel n
$\mathbf{w}[t]$	Matrice de décorrélation dans le domaine temporel à l'indice temporel t
$\mathbf{W}^l[k]$	Matrice de séparation pour l'indice fréquentiel k à la trame l
$(x_{clean})_u[t]$	Signal sans bruit utilisé pour l'entraînement à l'indice temporel t du locuteur u
$(x_{clean})_u^l[n]$	Signal sans bruit utilisé pour l'entraînement à l'indice temporel n du locuteur u à la trame l
$\mathbf{x}_{mic}[t]$	Ensemble des signaux à l'indice temporel t pour tous les microphones
$(x_{mic})_m[t]$	Échantillon à l'indice temporel t du microphone m
$(x_{mic})_m^l[n]$	Échantillon à l'indice temporel n du microphone m au sein de la trame l
$\mathbf{x}_{noisy}[t]$	Ensemble des signaux perçus par les microphones à l'indice temporel t
$(x_{noisy})_m[t]$	Signal perçu par le microphone m à l'indice temporel t
$x_{sep}[t]$	Processus aléatoire du signal séparé avec un indice temporel t

Variable	Définition
$x_{speech}[t]$	Processus aléatoire de la parole avec un indice temporel t
$(X_{clean})_u^l[k]$	Spectre sans bruit utilisé pour l'entraînement à l'indice fréquentiel k du locuteur u à la trame l
$(\mathbf{X}_{mic})^l[k]$	Ensemble des échantillons à l'indice fréquentiel k de tous les microphones au sein de la trame l
$(X_{mic})_m^l[k]$	Échantillon à l'indice fréquentiel k du microphone m au sein de la trame l
$ X_{post}(j\omega) ^2$	Densité spectrale de puissance du signal post-filtré à l'indice fréquentiel ω
$(X_{post})_s^l[k]$	Échantillon à l'indice fréquentiel k de la source post-filtrée s au sein de la trame l
$ X_{sep}(j\omega) ^2$	Densité spectrale de puissance du signal séparé à l'indice fréquentiel ω
$(\mathbf{X}_{sep})^l[k]$	Ensemble des échantillons à l'indice fréquentiel k de toutes les sources séparées au sein de la trame l
$(X_{sep})_s^l[k]$	Échantillon à l'indice fréquentiel k de la source séparée s au sein de la trame l
$(X_{speaker}^{beamformer})_s^l[k]$	Énergie du signal séparé à l'indice fréquentiel k pour la source séparée s au sein de la trame l
$(X_{speaker}^{powerMean})_s^l[k]$	Somme de l'énergie de tous les signaux des microphones à l'indice fréquentiel k pour la source séparée s au sein de la trame l
$(X_{speaker}^{signalMean})_s^l[k]$	Énergie de la somme de tous les signaux des microphones à l'indice fréquentiel k pour la source séparée s au sein de la trame l
$ X_{speech}(j\omega) ^2$	Densité spectrale de puissance de la parole à l'indice fréquentiel ω
$(X_{test}^{sep})_r^l[k]$	Puissance du spectre séparé utilisé pour le test à l'indice fréquentiel k du segment r à la trame l
$(X_{test}^{post})_r^l[k]$	Puissance du spectre post-filtré utilisé pour le test à l'indice fréquentiel k du segment r à la trame l
$(\{\mathbf{X}_{test}^{sep}\}_{emph})_r$	Puissance du spectre séparé avec emphase sur les hautes fréquences utilisé pour le test du segment r pour tous les filtres et les trames
$(\{X_{test}^{sep}\}_{emph})_r^l[k]$	Puissance du spectre séparé avec emphase sur les hautes fréquences à l'indice fréquentiel k pour le segment r à la trame l
$(\{\mathbf{X}_{test}^{post}\}_{emph})_r$	Puissance du spectre post-filtré avec emphase sur les hautes fréquences utilisé pour le test du segment r pour tous les filtres et les trames
$(\{X_{test}^{post}\}_{emph})_r^l[k]$	Puissance du spectre post-filtré avec emphase sur les hautes fréquences à l'indice fréquentiel k pour le segment r à la trame l

Variable	Définition
$(\{X_{test}^{sep}\}_{mel})_r^l[\Lambda]$	Puissance du spectre séparé associée au filtre Λ pour le segment r à la trame l
$(\{X_{test}^{post}\}_{mel})_r^l[\Lambda]$	Puissance du spectre post-filtré associée au filtre Λ pour le segment r à la trame l
$(\{X_{train}\}_{mel})_u^l[\Lambda]$	Puissance du spectre associée au filtre Λ du locuteur u à la trame l
$(X_{train}^{all})_u^l[k]$	Puissance du spectre sans bruit utilisé pour l'entraînement à l'indice fréquentiel k du locuteur u à la trame l
$(\{X_{train}^{all}\}_{emph})_u^l[k]$	Puissance du spectre filtré à l'indice fréquentiel k pour le locuteur u à la trame l
$(\hat{X}_{conv})_r[\Lambda]$	Moyenne des segments de parole au filtre Λ pour le segment r
$(\hat{\mathbf{X}}_{\text{conv}})_r$	Moyenne des segments de parole pour tous les filtres pour le segment r
$(\mathbf{z}_{\text{mic}})_m$	Vecteur de position du microphone m
$[(z_{mic})_m]_x$	Coordonnée en x de la position du microphone m
$[(z_{mic})_m]_y$	Coordonnée en y de la position du microphone m
$[(z_{mic})_m]_z$	Coordonnée en z de la position du microphone m
$(\mathbf{z}_{\text{part}}^{\text{new}})_s^l(f)$	Nouvelle position d'une particule f du filtre s à la trame l suite à l'échantillonnage
$(\mathbf{z}_{\text{part}})_s^l(f)$	Position de la particule f du filtre s au sein de la trame l
$[(z_{part})_s^l(f)]_x$	Coordonnée en x de la position de la particule f du filtre s au sein de la trame l
$[(z_{part})_s^l(f)]_y$	Coordonnée en y de la position de la particule f du filtre s au sein de la trame l
$[(z_{part})_s^l(f)]_z$	Coordonnée en z de la position de la particule f du filtre s au sein de la trame l
$(\mathbf{z}_{\text{part}}^{\text{new}})_s^l(f)$	Nouvelle position de la particule f du filtre s au sein de la trame l après échantillonnage
$(\mathbf{z}_{\text{part}}^{\cdot})_s^l(f)$	Vitesse de la particule f du filtre s au sein de la trame l
$[(z_{part}^{\cdot})_s^l(f)]_x$	Coordonnée en x de la vitesse de la particule f du filtre s au sein de la trame l
$[(z_{part}^{\cdot})_s^l(f)]_y$	Coordonnée en y de la vitesse de la particule f du filtre s au sein de la trame l
$[(z_{part}^{\cdot})_s^l(f)]_z$	Coordonnée en z de la vitesse de la particule f du filtre s au sein de la trame l
$(\mathbf{z}_{\text{part}'}^{\cdot})_s^l(f)$	Position non normalisée de la particule f du filtre s au sein de la trame l
$(\mathbf{z}_{\text{part}'}^{\cdot})_s^l(f)$	Vitesse non normalisée de la particule f du filtre s au sein de la trame l

Variable	Définition
$(\mathbf{z}_{\text{pot}})_q^l$	Position de la source potentielle q pour la trame l
$(\mathbf{z}_{\text{sph}})_d^{level}$	Coordonnées du point d sur la sphère unitaire au niveau $level$
$[(z_{sph})_d^{level}]_x$	Coordonnée en x du point d sur la sphère unitaire au niveau $level$
$[(z_{sph})_d^{level}]_y$	Coordonnée en y du point d sur la sphère unitaire au niveau $level$
$[(z_{sph})_d^{level}]_z$	Coordonnée en z du point d sur la sphère unitaire au niveau $level$
$(\mathbf{z}_{\text{track}})_s^l$	Position estimée de la source s à la trame l
$Z_s^l[k]$	Spectre filtré à l'indice fréquentiel k de la source séparée s à la trame l
$\alpha^l[k]$	Facteur de normalisation de l'énergie pour la séparation à l'indice fréquentiel k et la trame l
α_d	Taux d'adaptation pour l'estimation du rapport signal sur bruit
α_{emph}	Coefficient du filtre passe-haut
α_{pmin}	Taux d'adaptation minimal pour le post-filtrage
$\alpha_s^l(f)$	Paramètre qui contrôle le paramètre d'amortissement associé à une particule f au sein d'un filtre s pour la trame l
α_s	Taux d'adaptation pour l'estimation du bruit stationnaire
$(\alpha_{sep}^p)_s^l[k]$	Taux d'adaptation pour le post-filtrage à l'indice fréquentiel k de la source s à la trame l
α_{score}	Seuil de la fonction sigmoïde pour l'indice de confiance
α_t	Taux pour lisser le spectre des sources séparées pour le post-filtrage
α_u	Facteur de mises à jour du bruit
α_ζ	Taux d'adaptation pour la moyenne récursive pour le rapport signal sur bruit <i>a priori</i> estimé
$\beta_s^l(f)$	Paramètre qui contrôle le paramètre d'excitation associé à une particule f au sein d'un filtre s pour la trame l
β_{score}	Inverse de la pente de la fonction sigmoïde pour l'indice de confiance
γ	Paramètre représentant le temps de réverbération
$(\gamma_{sep})_s^l[k]$	Rapport signal sur bruit <i>a posteriori</i> à l'indice fréquentiel k de la source s de la trame l
δ	Rapport signal sur réverbération
$(\Delta_{\text{sph}})_v^{level}$	Ensemble des sommets du triangle v de la sphère unitaire au niveau $level$
$[(\Delta_{\text{sph}})_v^{level}]_1$	Premier sommet du triangle v de la sphère unitaire au niveau $level$

Variable	Définition
$([(\Delta_{sph})_v^{level}]_1)_x$	Coordonnée en x du premier sommet du triangle v de la sphère unitaire au niveau $level$
$([(\Delta_{sph})_v^{level}]_1)_y$	Coordonnée en y du premier sommet du triangle v de la sphère unitaire au niveau $level$
$([(\Delta_{sph})_v^{level}]_1)_z$	Coordonnée en z du premier sommet du triangle v de la sphère unitaire au niveau $level$
$[(\Delta_{sph})_v^{level}]_2$	Deuxième sommet du triangle v de la sphère unitaire au niveau $level$
$([(\Delta_{sph})_v^{level}]_2)_x$	Coordonnée en x du deuxième sommet du triangle v de la sphère unitaire au niveau $level$
$([(\Delta_{sph})_v^{level}]_2)_y$	Coordonnée en y du deuxième sommet du triangle v de la sphère unitaire au niveau $level$
$([(\Delta_{sph})_v^{level}]_2)_z$	Coordonnée en z du deuxième sommet du triangle v de la sphère unitaire au niveau $level$
$[(\Delta_{sph})_v^{level}]_3$	Troisième sommet du triangle v de la sphère unitaire au niveau $level$
$([(\Delta_{sph})_v^{level}]_3)_x$	Coordonnée en x du troisième sommet du triangle v de la sphère unitaire au niveau $level$
$([(\Delta_{sph})_v^{level}]_3)_y$	Coordonnée en y du troisième sommet du triangle v de la sphère unitaire au niveau $level$
$([(\Delta_{sph})_v^{level}]_3)_z$	Coordonnée en z du troisième sommet du triangle v de la sphère unitaire au niveau $level$
$[(\Delta_{sph})_v^{level}]_A$	Premier sommet intermédiaire du triangle v de la sphère unitaire au niveau $level$
$[(\Delta_{sph})_v^{level}]_B$	Deuxième sommet intermédiaire du triangle v de la sphère unitaire au niveau $level$
$[(\Delta_{sph})_v^{level}]_C$	Troisième sommet intermédiaire du triangle v de la sphère unitaire au niveau $level$
Δ_{score}	Écart entre les deux pointages les plus faibles
ΔT	Intervalle de temps entre les mises à jour des particules
ϵ_{log}	Seuil minimum pour éviter une énergie nulle pour un filtre
$\zeta_m^l[k]$	Facteur de pondération à l'indice fréquentiel k du microphone m au sein de la trame l
$(\zeta_{frame})_s^l[k]$	Moyenne récursive pour le rapport signal sur bruit <i>a priori</i> sur une fenêtre entière estimé à l'indice fréquentiel k , la source s et la trame l
$(\zeta_{frame})_u^l[k]$	Moyenne récursive pour le rapport signal sur bruit <i>a priori</i> sur une fenêtre entière estimé à l'indice fréquentiel k , le locuteur u et la trame l

Variable	Définition
$(\zeta_{global})_s^l[k]$	Moyenne récursive pour le rapport signal sur bruit <i>a priori</i> sur une fenêtre large estimé à l'indice fréquentiel k , la source s et la trame l
$(\zeta_{global})_u^l[k]$	Moyenne récursive pour le rapport signal sur bruit <i>a priori</i> sur une fenêtre large estimé à l'indice fréquentiel k , le locuteur u et la trame l
$(\zeta_{local})_s^l[k]$	Moyenne récursive pour le rapport signal sur bruit <i>a priori</i> sur une fenêtre étroite estimé à l'indice fréquentiel k , la source s et la trame l
$(\zeta_{local})_u^l[k]$	Moyenne récursive pour le rapport signal sur bruit <i>a priori</i> sur une fenêtre étroite estimé à l'indice fréquentiel k , le locuteur u et la trame l
η	Gain pour diminuer l'amplitude des sources séparées durant le post-filtrage
θ	Seuil d'activité vocale pour le post-filtrage
θ_{scale}	Facteur de mise à l'échelle pour la détection de l'activité vocale
θ_u	Seuil d'activité vocale pour la reconnaissance selon le locuteur u
κ	Vecteur pour évaluer rapidement le gain transcendant
$(\kappa_{ac})_u^v[\Lambda]$	Centroïde v normalisé du locuteur u au filtre Λ
$(\kappa_{noisy})_u^v[\Lambda]$	Centroïde v mis à jour du locuteur u au filtre Λ
$\kappa_s^l(f)$	Réalisation d'un processus aléatoire blanc avec une distribution uniforme entre 0 et 1 associé à une particule f d'un filtre s pour une trame l
κ_u^v	Centroïde v du locuteur u pour tous les filtres
$\kappa_u^v[\Lambda]$	Centroïde v du locuteur u au filtre Λ
λ	Taux de régularisation pour la séparation
$(\lambda_{mic})_m^l[k]$	Bruit total à l'indice fréquentiel k du microphone m au sein de la trame l
$(\lambda_{mic}^{rev})_m^l[k]$	Réverbération à l'indice fréquentiel k du microphone m au sein de la trame l
$(\lambda_{mic}^{stat})_m^l[k]$	Bruit stationnaire à l'indice fréquentiel k du microphone m au sein de la trame l
$[(\lambda_{mic}^{stat})_f]_m^l[k]$	Spectre filtré dans le domaine fréquentiel à l'indice fréquentiel k du microphone m au sein de la trame l
$[(\lambda_{mic}^{stat})_{min}]_m^l[k]$	Bruit stationnaire minimum à l'indice fréquentiel k du microphone m au sein de la trame l

Variable	Définition
$[(\lambda_{mic}^{stat})_t]_m^l[k]$	Spectre filtré dans le domaine temporel à l'indice fréquentiel k du microphone m au sein de la trame l
$[(\lambda_{mic}^{stat})_{tmp}]_m^l[k]$	Bruit stationnaire temporaire à l'indice fréquentiel k du microphone m au sein de la trame l
$(\lambda_{sep})_s^l[k]$	Bruit total à l'indice fréquentiel k de la source s au sein de la trame l
$(\lambda_{sep}^{leak})_s^l[k]$	Bruit à l'indice fréquentiel k provenant des sources séparées autre que la source s à la trame l
$(\lambda_{sep}^{stat})_s^l[k]$	Bruit stationnaire à l'indice fréquentiel k de la source séparée s au sein de la trame l
Λ	Indice d'un filtre dans le banc de filtres
Λ_{max}	Nombre de filtres dans le banc de filtres
μ	Taux d'adaptation pour la séparation
$(\xi_{mic})_m^l[k]$	Estimation du rapport signal sur bruit <i>a priori</i> à l'indice fréquentiel k du microphone m au sein de la trame l
$(\xi_{sep})_s^l[k]$	Rapport signal sur bruit <i>a priori</i> à l'indice fréquentiel k de la source s de la trame l
τ	Délai en échantillons
$\tau_{m,s}^l$	Délai d'arrivée pour la séparation pour le microphone m , la source s et la trame l
τ_{max}	Délai d'arrivée du signal entre les microphones m_1 et m_2 correspondant au point sur la sphère possédant l'énergie maximale pour la source potentielle q à la trame l
τ_{range}	Étendue de la plage de remise à zéro durant la recherche sur la sphère unitaire
v	Indice d'un triangle de la sphère unitaire
Υ_{level}	Nombre de triangles au niveau <i>level</i> sur la sphère unitaire
$(\phi_{int})_s^l[k]$	Multiplicateur entier pour le calcul du gain transcendant à l'indice fréquentiel k de la source s à la trame l
$(\phi_{frac})_s^l[k]$	Multiplicateur fractionnaire pour le calcul du gain transcendant à l'indice fréquentiel k de la source s à la trame l
$(\chi_{int})_s^l[k]$	Partie entière pour le calcul du gain transcendant à l'indice fréquentiel k de la source s à la trame l
$(\chi_{frac})_s^l[k]$	Partie fractionnaire pour le calcul du gain transcendant à l'indice fréquentiel k de la source s à la trame l
Ψ_s^l	Variable aléatoire discrète dont la densité de probabilité est reliée à la pondération des particules du filtre s à la trame l

Variable	Définition
ω_{frame}	Largeur de la fenêtre entière
ω_{global}	Largeur de la fenêtre large
ω_{local}	Largeur de la fenêtre étroite
$\omega_s^l(f)$	Pondération de chaque particule f pour la source s à la trame l
Ω	Différence entre le modèle et les caractéristiques vocales

CHAPITRE 1

INTRODUCTION

Le vieillissement de la population est un phénomène qui affectera grandement les pays industrialisés au cours des prochaines décennies. En effet, on estime qu'en 2050, environ 80 millions d'Américains (c'est-à-dire 20% de la population à ce moment) seront âgés de 65 ans et plus [20]. De plus, avec la pénurie de personnel médical, il sera difficile de fournir des soins de qualité aux personnes situées en région éloignée. Dans un tel contexte, les robots mobiles et autonomes devraient être de plus en plus mis à contribution afin de fournir des soins à distance à ces personnes [10]. Les médecins pourraient ainsi faire un suivi de l'état de santé du patient et effectuer certaines manipulations de base. Bien évidemment, un robot ne peut pas être contrôlé à distance par un opérateur de façon continue. Lorsque le robot est laissé à lui-même, il doit pouvoir interagir avec les personnes d'une façon autonome. Entre autres, cette autonomie devra inclure une interaction grâce à la voix humaine [29].

Le contexte actuel démontre les besoins grandissants pour une audition artificielle en robotique mobile. Les techniques de localisation, suivi et séparation de sources sonores ont déjà fait leurs preuves dans un environnement bruité. Cependant, ces techniques commencent seulement à faire leur apparition en robotique mobile depuis quelques années. En effet, l'augmentation récente de la puissance de calcul des processeurs permet dorénavant d'intégrer ces techniques complexes de reconnaissance à des systèmes embarqués en temps réel. Par exemple, le robot SIG est en mesure de séparer jusqu'à trois sources sonores et d'effectuer une reconnaissance de la parole sur chacune d'entre elles [27]. Pour sa part, le robot SDR-4x de Sony peut reconnaître plusieurs individus grâce à une combinaison de leurs caractéristiques vocales et faciales [14]. D'autre part, les robots Spartacus, SIG2 et ASIMO utilisent le système ManyEars pour effectuer une reconnaissance de la parole avec un vocabulaire restreint à partir de plusieurs sources sonores séparées simultanément [32], [52], [53].

L'audition artificielle en robotique vise à effectuer une reconnaissance de la parole pour permettre au robot d'interagir avec plusieurs utilisateurs. Cependant, le domaine de la reconnaissance de locuteurs dans un environnement bruité est également un sujet de recherche populaire. Ce genre de système est particulièrement utile dans le contexte de la robotique mobile car il permettrait au robot de reconnaître l'identité de ou des interlocuteurs à partir

de ses caractéristiques vocales. De plus, l'identification vocale de la personne pourrait être combinée à d'autres informations (la perception visuelle par exemple) pour augmenter la robustesse au niveau de l'identification des individus.

Ce mémoire vise à démontrer qu'il est possible de concevoir un système de reconnaissance de locuteurs pour un robot mobile qui permet d'associer l'identité d'une personne avec sa position physique dans une pièce par rapport au robot. Le système proposé permet l'identification de locuteurs dans un contexte où un seul locuteur principal parle à la fois, avec certains chevauchements de locuteurs sporadiques comme c'est souvent le cas dans une conversation naturelle dans un groupe d'individus. Le système de reconnaissance de locuteurs réalisé, baptisé WISS (*Who IS Speaking*), est couplé au système ManyEars initialement conçu pour être utilisé avec un système de reconnaissance de la parole. Ce dernier effectue la localisation, le suivi, la séparation et le post-filtrage des sources sonores, et s'avère une solution attrayante en raison du compromis intéressant qu'il offre entre les performances et l'utilisation en temps réel et son exploitation sur plusieurs robots [32], [52], [53]. En accédant directement à certains signaux propres au traitement réalisé par ManyEars, il est ainsi possible d'ajouter une capacité de reconnaissance de locuteurs à un robot mobile avec une très faible augmentation en coût de calculs. Cette stratégie est particulièrement intéressante puisque l'audition artificielle en robotique doit s'effectuer en temps réel et que les ressources matérielles pour effectuer le traitement sur un robot mobile sont limitées en raison des contraintes physiques (dimensions, poids), énergétiques et de calculs disponibles sur le robot.

Le mémoire est organisé de la façon suivante. Le chapitre 2 traite de l'état de l'art en ce qui concerne la reconnaissance de locuteurs en robotique. Par la suite, le chapitre 3 décrit le système ManyEars utilisé pour effectuer la localisation, le suivi, la séparation et le post-filtrage. Le chapitre 4 expose le système de reconnaissance de locuteurs WISS. Finalement, le chapitre 5 présente les résultats obtenus dans des scénarios de tests mettant en évidence les capacités de WISS dans un environnement bruité.

CHAPITRE 2

RECONNAISSANCE DE LOCUTEURS EN ROBOTIQUE

D'une manière générale, l'utilisation de la voix humaine dans un contexte de biométrie se divise en deux grandes catégories : l'identification et la vérification du locuteur (réf. : chapitre 18 [23]). Pour l'identification, le locuteur inconnu est comparé à un ensemble fini de locuteurs connus. Un pointage est retourné en fonction des similitudes entre le locuteur inconnu et chaque élément de l'ensemble. Le locuteur de l'ensemble qui possède la plus grande similitude avec le locuteur inconnu est retenu. Il est alors déterminé que le locuteur inconnu correspond au locuteur identifié au sein de l'ensemble. Bien entendu, la plus grande faiblesse de ce système concerne l'apparition d'un nouveau locuteur qui ne figure pas dans l'ensemble préétabli. Dans ce cas, le système associe au nouveau locuteur l'identité d'un locuteur de l'ensemble, ce qui entraîne forcément une mauvaise identification. En contrepartie, la vérification du locuteur permet de palier au problème précédent en ajoutant une hypothèse selon laquelle le locuteur inconnu n'appartient pas nécessairement à l'ensemble des locuteurs connus. Cette méthode est généralement utilisée à des fins d'authentification : le locuteur indique son identité, que le système valide par biométrie. En effet, si le locuteur ne présente pas suffisamment de similitudes avec un des locuteurs connus, ce dernier est perçu comme un imposteur. Il est donc possible d'utiliser ce système à des fins de sécurité pour autoriser ou refuser un accès à partir de la voix du locuteur. D'autre part, il est possible d'associer un imposteur à un locuteur connu (*false acceptance*) ou de méprendre un locuteur connu pour un imposteur (*false rejection*). En général, l'utilisation d'un modèle universel pour représenter l'ensemble des locuteurs inconnus permet de comparer chaque locuteur à un modèle générique afin d'améliorer les performances [38].

Chaque locuteur est unique à différents niveaux. Le premier niveau regroupe les propriétés acoustiques des phonèmes du locuteur. Par la suite viennent le rythme de la parole, l'intonation dans la voix et les changements de volume. Finalement, la sémantique, le vocabulaire utilisé, la prononciation, etc. permettent également d'identifier le locuteur [40]. Les phonèmes sont générées par une cascade de deux composantes : une source d'excitation et un conduit vocal. La source d'excitation est composée d'une série d'impulsions à une fréquence précise, qui représentent les vibrations des cordes vocales lorsque des sons voisés sont produits. Cette fréquence définit le ton de voix du locuteur. Les voyelles, les con-

sonnes vibrantes, latérales, nasales ainsi que certaines consonnes occlusives et fricatives sont dites voisées. Lorsqu'il s'agit de sons non voisés, la source d'excitation est remplacée par un générateur de bruit blanc. D'autres consonnes occlusives et fricatives sont dites non voisées. Par exemple, dans le mot "silence", la consonne fricative [s] est dit non voisée, tandis que la voyelle [i] et la consonne latérale [l] sont dites voisées. Le signal produit par cette source est ensuite modifié par un filtre qui représente le conduit vocal. Cette section peut être modélisée comme une chaîne de tubes connectés en série [7]. La forme unique de ces tubes génère une réponse fréquentielle propre à chaque individu. La figure 2.1 illustre ce modèle. Les propriétés acoustiques des phonèmes sont couramment utilisées pour la reconnaissance de locuteurs. Récemment, la recherche a été accentuée sur les caractéristiques plus complexes, tel que le choix des phonèmes par chaque locuteur [36], [56]. Cette méthode est prometteuse mais requiert un modèle général préentraîné pour chaque phonème. Ceci implique entre autres que la langue parlée par le locuteur soit connue à l'avance.

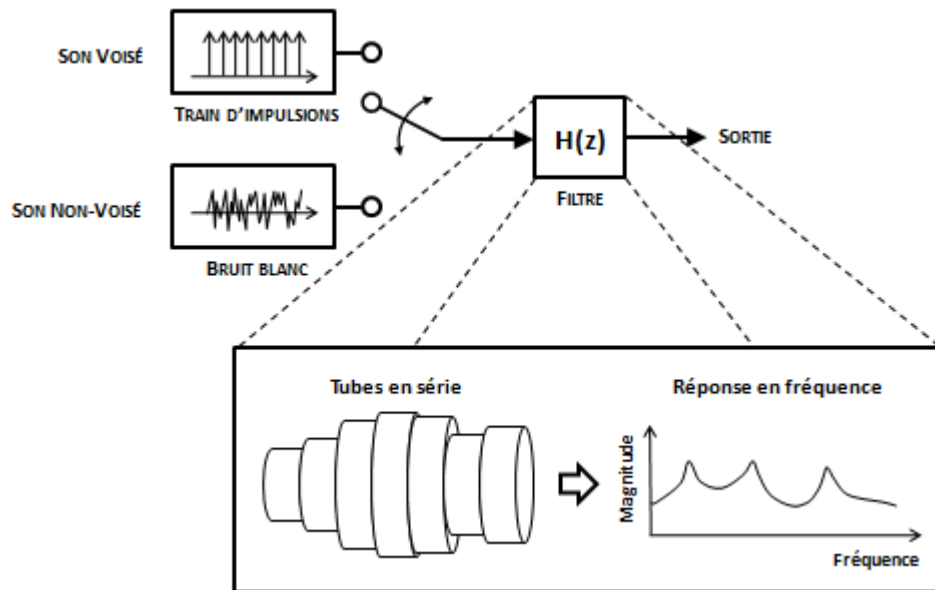


Figure 2.1 Modèle de la parole

Plusieurs caractéristiques vocales peuvent être utilisées pour représenter les phonèmes. Les plus utilisées sont les coefficients cepstraux répartis sur l'échelle Mel (*Mel Frequency Cepstral Coefficients* (MFCC)) [42]. Les coefficients à prédiction linéaire (*Linear Predictive Coefficients* (LPC)) ont également démontré de bonnes performances (réf. : chapitre 6 [23]). De plus, il est possible d'utiliser des caractéristiques similaires aux MFCC mais dans le domaine spectral [53]. Ceci permet l'utilisation de masques pour ignorer les bandes bruitées [43].

Les caractéristiques vocales peuvent être modélisées de différentes manières. La quantification vectorielle (*Vector Quantization* (VQ)) est une technique simple et efficace qui permet de représenter l'ensemble des caractéristiques d'un locuteur par un dictionnaire (livre de code). La technique k-moyennes (*k-means*) est généralement utilisée pour entraîner les modèles [6]. Il est également possible de représenter la distribution probabiliste des caractéristiques de chaque individu par une somme pondérée de distributions gaussiennes. Ce modèle, le modèle de mélange gaussien (*Gaussian Mixture Model* (GMM)), est entraîné à l'aide de l'algorithme de maximisation de vraisemblance (*maximum likelihood*). Il offre généralement des performances supérieures à celles du VQ. Il nécessite toutefois plus de calculs durant l'entraînement et la reconnaissance [41].

En général, les performances pour la reconnaissance de locuteurs sont satisfaisantes lorsque la voix du locuteur est à proximité du microphone [39], mais elles se dégradent rapidement lorsque la reconnaissance s'effectue par l'entremise d'un microphone éloigné [1], [19], [57]. Cette dégradation est causée par deux types de bruits : le bruit convolutif et le bruit additif.

Plusieurs stratégies ont été proposées pour réduire les effets du bruit convolutif. La technique de normalisation des moyennes cepstrales (*Cepstral Mean Normalization* (CMN)) et ses versions améliorées sont devenues un incontournable pour palier à ce problème [16], [22]. Il s'agit de soustraire la moyenne du cepstre qui comprend principalement les caractéristiques du canal pour conserver uniquement les caractéristiques du locuteur. Il est également possible d'effectuer cette opération à l'aide d'un filtre passe-haut, ce qui facilite la normalisation du canal en temps réel (réf. : chapitre 10 [23]). Une autre technique consiste à évaluer le facteur de CMN en fonction de la position du locuteur [48]. L'inconvénient de cette méthode est que le facteur de normalisation doit être calculé *a priori* pour chaque position de l'environnement. En conséquence, cette technique n'est pas appropriée pour un environnement dynamique. En général, il est difficile de normaliser le canal lorsqu'il y a un court segment de parole. En effet, pour ce cas particulier, la moyenne du cepstre contient des caractéristiques propres au locuteur.

Le bruit additif affecte également les performances de la reconnaissance de locuteurs. Il est possible d'en diminuer l'impact en effectuant une soustraction spectrale [9], [34], [47]. Pour sa part, au lieu de chercher à normaliser les caractéristiques, la combinaison de modèle parallèle (*Parallel Model Combination* (PMC)) permet d'intégrer le bruit additif au modèle précédemment entraîné dans des conditions idéales [51]. La normalisation se fait donc au niveau du modèle et ce dernier est comparé avec les caractéristiques bruitées.

Dans notre projet, le grand défi consiste donc à effectuer une reconnaissance de locuteurs dans un environnement qui contient du bruit convolutif et du bruit additif. C'est le cas de la plupart des scénarios lorsque le microphone est éloigné du locuteur. Il est possible de réduire le bruit additif et ensuite réduire le bruit convolutif présent dans les caractéristiques. Cette normalisation en cascade est démontré dans [21]. Bien que cette technique améliore les performances en général, son effet reste toutefois limité lorsque le rapport signal sur bruit (*Signal-to-Noise Ratio* (SNR)) est faible. Pour la reconnaissance de la parole (et non de locuteurs), il est possible d'adapter les modèles de Markov cachés (*Hidden Markov Models* (HMM)) pour qu'ils correspondent à l'environnement bruité [18]. Bien qu'elle soit conçue pour la reconnaissance de la parole, cette technique présente des éléments intéressants qui pourraient être utilisés pour la reconnaissance de locuteurs. Cependant, la complexité des calculs rend difficile l'utilisation de cet algorithme en temps réel.

L'intérêt pour la reconnaissance de locuteurs dans le cadre d'une application robotique n'est pas nouveau. Par exemple, en utilisant les caractéristiques vocales d'un individu, il est possible de classifier les locuteurs en fonction de leurs traits significatifs tels que leur âge et leur genre [54]. Ce système suppose que l'interaction vocale se fait dans un environnement qui possède une réverbération importante (comme dans une maison par exemple) et que les sources sonores sont en mouvement. À partir de l'entrée audio de trois microphones, différentes familles de caractéristiques vocales sont extraites. L'algorithme sélectionne automatiquement les caractéristiques qui offrent la meilleure discrimination pour classifier les locuteurs. Il est observé que le choix des caractéristiques offrant une meilleure discrimination est différent selon l'environnement dans lequel évolue le robot. Les résultats démontrent que les caractéristiques MFCC offrent à elles seules un taux de reconnaissance moyen de 92.45% pour tous les environnements confondus. Par contre, en incluant d'autres familles de caractéristiques qui offrent des performances supérieures pour certains environnements, ce taux de reconnaissance augmente à 93.51%. Bien qu'un faible gain soit présent, cette expérience suggère que les caractéristiques MFCC permettent de discriminer efficacement les locuteurs.

Il est également possible d'identifier un individu en combinant ses caractéristiques vocales et ses traits faciaux. Dans le travail de Ban [4], un modèle de mixture gaussienne et les caractéristiques MFCC sont utilisés pour la reconnaissance de locuteurs. La reconnaissance faciale est effectuée à l'aide de la méthode *fisherface*, moins sensible aux différences d'éclairage et aux expressions faciales. Une caméra et un seul microphone sont utilisés pour cette application. La reconnaissance faciale offre un taux de reconnaissance de 99.5%

lorsque la personne se situe à une distance d'un mètre du robot, mais les performances se dégradent rapidement à une distance de 3 mètres pour atteindre un taux de 43.25%. Pour sa part, à une distance d'un mètre, la reconnaissance de locuteur offre un taux de reconnaissance de 88.25%, ce qui est nettement inférieur au taux obtenu pour la reconnaissance faciale. Cependant, à une distance de trois mètres, ce taux diminue seulement à 86.5%. La combinaison de ces deux techniques mènent jusqu'à des taux de reconnaissance de 99.5% pour une distance d'un mètre et 88.25% pour une distance de trois mètres. Ces résultats démontrent entre autres qu'une reconnaissance vocale de locuteurs est plus robuste au changement de position d'une personne par rapport à une reconnaissance faciale. Cette expérience a cependant été menée dans un environnement peu bruyé, ce qui correspond rarement aux scénarios typiques lors d'une interaction humain-robot.

Un système de reconnaissance de locuteurs à huit microphones est également proposé [24]. Ce système effectue une identification à partir du signal de chaque microphone et détermine une décision globale qui découle de l'ensemble de ces décisions individuelles. Le taux de bonnes identifications se situent à environ 90% lorsque le locuteur se situe à une distance de deux mètres et diminue à 75% à une distance de trois mètres. Le SNR n'est pas indiqué mais les expériences sont effectuées dans une pièce silencieuse.

CHAPITRE 3

MANYEARS, UN SYSTÈME DE LOCALISATION, SUIVI ET SÉPARATION

Plusieurs techniques ont été proposées dernièrement pour permettre de localiser avec précision une ou plusieurs sources sonores. C'est le cas de la référence [44] qui propose un système qui s'inspire du pavillon auriculaire de l'être humain en utilisant deux microphones. Grâce à la différence de phase entre les signaux obtenus, ce système est en mesure d'établir la position angulaire d'une seule source sonore. Ce type de système ajuste la différence de phase en fonction de chaque fréquence présente dans le spectre sonore obtenu [25].

Contrairement au système anthropomorphe précédent, le système ManyEars est composé de huit microphones omnidirectionnels. Ce système a démontré qu'il est possible pour un robot mobile de localiser et de suivre plusieurs sources sonores simultanément [45]. La figure 3.1 illustre le traitement interne des modules de ManyEars. La localisation s'effectue grâce à un formateur de faisceaux (*beamformer*) qui analyse les signaux sonores provenant de chaque microphone. Pour cette application, la position angulaire des individus est obtenue mais la distance entre chaque personne et le robot est ignorée. Afin d'augmenter les performances du système, la densité spectrale du bruit ambiant est estimée et une pondération des bandes fréquentielles est effectuée. Malgré les différentes améliorations proposées, le formateur de faisceaux fournit tout de même des informations bruitées en ce qui a trait à la position des sources sonores. C'est pour cette raison qu'un filtre particulière est utilisé afin d'effectuer un suivi intelligent de la position angulaire de chaque source sonore. Il est également possible d'utiliser un filtre de Kalman [17] mais le filtre particulière permet une meilleure gestion des fausses détections et de l'assignation des sources. Une étude comparative de différentes techniques de localisation fonctionnant en temps réel sur un processeur généraliste (*General Purpose Processor* (GPP)) confirme que celle utilisée par ManyEars (formateur de faisceaux avec une pondération des bandes et recherche sur une sphère avec une tessellation triangulaire) offre une précision supérieure aux techniques traditionnelles (formateur de faisceaux sans pondération des bandes et recherche sur une sphère avec une tessellation rectangulaire) [3]. ManyEars fut aussi implémenté sur un processeur de signal numérique (*Digital Signal Processor* (DSP)), qui consomme beaucoup moins de puissance qu'un GPP [5]. Pour cette implémentation sur un DSP TMS320C6713,

la localisation est cependant moins précise et le nombre de sources sonores pouvant être détectées simultanément est limité à deux afin de réduire la quantité de calculs nécessaires.

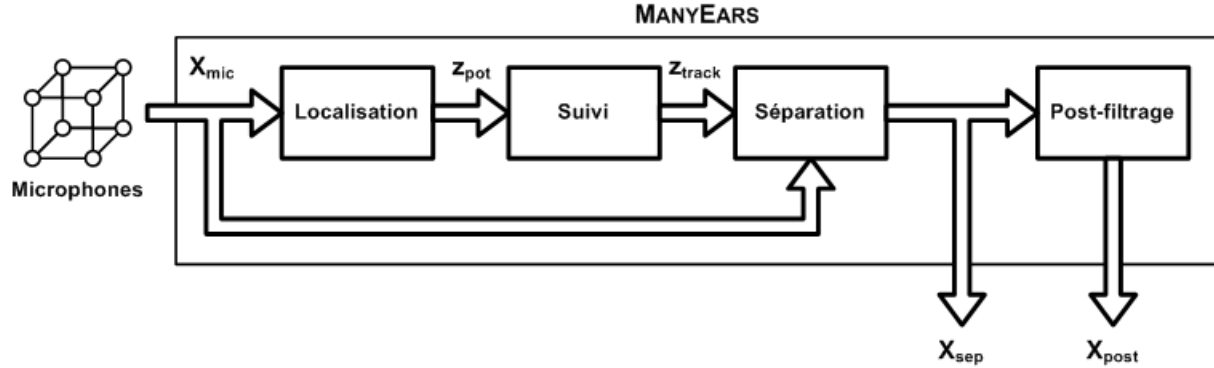


Figure 3.1 Schéma-bloc du système ManyEars

D'autre part, un formateur de faisceaux utilisant 128 microphones permet une meilleure résolution spatiale comparativement au système ManyEars qui utilise seulement huit microphones [13]. Le nombre élevé de microphones permet également de déterminer la distance entre chaque personne et le robot. Cependant, ce système peut difficilement fonctionner en temps réel vu la charge de calculs associée à une quantité élevée de microphones. En récoltant des données audio pendant une plus longue période de temps lorsque le robot se déplace dans son environnement, il serait possible de distinguer les sources de bruit stationnaires (comme les ventilateurs dans la pièce) des locuteurs en mouvement [31]. Le système pourrait alors cesser de localiser et suivre ces sources de bruits pour se concentrer uniquement sur les locuteurs dans la pièce.

Un être humain est en mesure de suivre une conversation dans un environnement bruité où plusieurs personnes discutent simultanément. La séparation des sources sonores est un problème complexe qui est communément appelé l'effet cocktail party (*cocktail party effect*). En supposant que la position de chaque source sonore est initialement connue, il est possible de séparer les sources sonores avec un éliminateur de lobes latéraux (*Generalized Sidelobe Canceller* (GSC)). Cette technique peut être mise en place à l'aide d'un filtre simple qui effectue une opération de délai et de sommation. Un lobe principal et plusieurs lobes secondaires sont ainsi formés de manière à amplifier le signal en direction de la source désirée. Cependant, les lobes secondaires introduisent du bruit en provenance des autres sources sonores et ainsi les résultats de séparation obtenus ne sont pas optimaux [28].

Comme autre solution, l'analyse en composantes indépendantes (*Independent Component Analysis* (ICA)) part du principe que les sources sonores sont représentées par des processus aléatoires non-stationnaires et indépendants. L'idée consiste à utiliser une matrice de

décorrélation qui minimise la corrélation croisée entre les sources sonores [30]. Cette technique permet de séparer statistiquement les sources sonores mais ne considère pas leurs positions respectives. En effet, la position de chaque source fournit une information supplémentaire qui permet une meilleure discrimination, en particulier dans un environnement réverbérant.

Il est possible de combiner les deux techniques précédentes afin de tirer avantage de la position des sources sonores et de leur indépendance statistique. Deux techniques sont proposées : la séparation géométrique des sources (*Geometric Source Separation* (GSS)) et la séparation avec un critère minimum d'information mutuelle (*Adaptive Beamforming with a Minimum Mutual Information Criterion* (ABMMIC)). Le GSS est une technique récursive qui permet de séparer le spectre fréquentiel de plusieurs sources sonores en réduisant l'interférence [37]. Il s'agit d'un algorithme récursif qui met à jour une matrice de séparation en fonction de plusieurs contraintes statistiques. Cette technique est d'autant plus intéressante qu'elle permet l'ajout et le retrait de sources sonores. La matrice de séparation est mise à jour en fonction de plusieurs paramètres définis. Ces paramètres sont habituellement déterminés expérimentalement. Cependant, il a été démontré qu'il est possible d'ajuster automatiquement ces paramètres en fonction de l'environnement dans lequel se trouve le robot. Ceci a pour effet d'optimiser la séparation des sources. Cette technique a d'ailleurs été utilisée en temps réel au sein du robot Honda ASIMO [35] en utilisant à la base le système ManyEars [45]. Le ABMMIC est une technique similaire à celle de la séparation géométrique, mais qui permet une adaptation rapide des termes de séparation lorsque plusieurs sources sont actives simultanément. De plus, les sources sonores sont considérées comme des processus aléatoires possédant une distribution gamma, K0 ou Laplacienne, ce qui se rapproche plus des phénomènes physiques observés. Cette nouvelle technique permet entre autres une réduction plus importante des interférences provenant de la réverbération de la pièce. Les performances obtenues pour la séparation avec un critère minimum d'information mutuelle sont légèrement supérieures à celles de la séparation géométrique de sources [28]. Cependant, le GSS requiert une complexité inférieure de calculs, ce qui en fait une technique plus appropriée dans le cadre d'un système en temps réel.

Une fois les sources sonores séparées, le signal associé à chaque source demeure affecté par des interférences provenant des autres sources sonores et de bruits stationnaires présents dans la pièce. C'est pour cette raison qu'un filtre est utilisé pour diminuer la pondération des bandes fréquentielles contaminées par ce bruit [45]. En diminuant le poids des bandes bruitées, il est possible d'augmenter l'importance des bandes non bruitées. Par la suite,

des composantes cepstrales sont extraites à partir de ces bandes dans le but d'effectuer une reconnaissance vocale de la parole et de locuteurs. Cependant, certaines composantes sont affectées par les bandes bruitées et ne sont pas corrélées statistiquement avec les composantes valides. Un masque est donc généré grâce aux propriétés statistiques de chaque composante et est appliqué afin de réduire la contribution des composantes erronées durant le processus de reconnaissance [46].

Les prochaines sections expliquent plus en détails le fonctionnement des modules de ManyEars. De plus, ManyEars exploite plusieurs paramètres internes fixés empiriquement [32], et ceux-ci sont identifiés et expliqués.

3.1 Localisation de sources sonores avec ManyEars

La première étape de traitement au sein du système ManyEars consiste à effectuer une localisation à l'aide d'un formateur de faisceaux afin de proposer des positions potentielles des locuteurs. Ce processus est illustré à la figure 3.2.

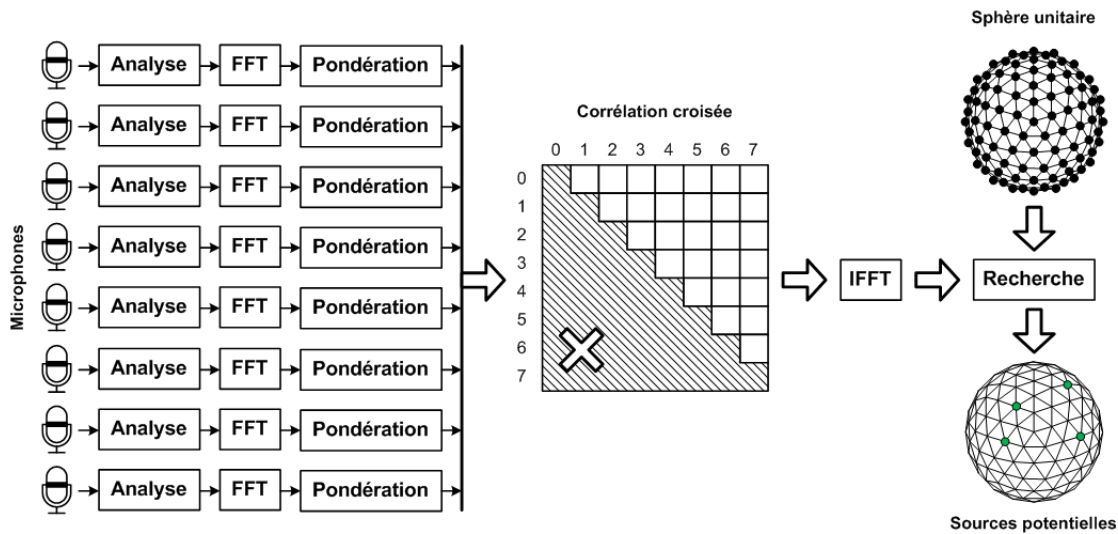


Figure 3.2 Aperçu de la localisation avec ManyEars

Le vecteur $\mathbf{x}_{\text{mic}}[t]$ représente l'ensemble des signaux $(x_{\text{mic}})_m[t]$ pour chaque échantillon à l'indice t à chaque microphone m pour un total de M microphones, tel que démontré dans l'équation 3.1.

$$\mathbf{x}_{\text{mic}}[t] = \begin{bmatrix} (x_{\text{mic}})_0[t] \\ (x_{\text{mic}})_1[t] \\ \dots \\ (x_{\text{mic}})_{M-1}[t] \end{bmatrix} \quad (3.1)$$

Les microphones sont disposés en cube et la position de chacun est définie par rapport au centre du cube, tel que défini dans l'équation 3.2.

$$(\mathbf{z}_{\text{mic}})_m = \begin{bmatrix} [(z_{\text{mic}})_m]_x \\ [(z_{\text{mic}})_m]_y \\ [(z_{\text{mic}})_m]_z \end{bmatrix} \quad 0 \leq m < M \quad (3.2)$$

3.1.1 Analyse et FFT

Le signal de chaque microphone est d'abord multiplié par une fenêtre définie par l'équation 3.3. Cette opération permet de réduire l'étalement spectral. Dans cette équation, $(x_{\text{mic}})_m^l[n]$ représente le signal de la trame l du microphone m à l'échantillon n . Chaque trame possède N échantillons ($N = 1024$ dans le cas présent) et il y a un total de L_{total} trames. Un chevauchement de 50% est effectué en fixant $o = 0.5$.

$$(x_{\text{mic}})_m^l[n] = w[n](x_{\text{mic}})_m[n + (l)(N)(o)] \quad 0 \leq n < N, \quad 0 \leq l < L_{\text{total}} \quad (3.3)$$

La fenêtre $w[n]$ utilisée n'est pas une fenêtre de Hann ou Hamming comme c'est souvent le cas pour ce type d'application. En effet, le signal original est d'abord multiplié par la fenêtre $w[n]$. Chaque trame est ensuite multipliée à nouveau par cette même fenêtre dans l'étape de synthèse (optionnelle suite à la séparation des sources mais permettant de récupérer les signaux séparés dans le domaine temporel à des fins d'écoute) et ensuite additionnée et chevauchée. Suite à ces opérations, le signal final doit conserver un gain unitaire constant, pour éviter un battement causé par un gain variable dans le temps. Pour cette raison, la fenêtre doit satisfaire la propriété de l'équation 3.4. La fenêtre $w[n]$ de l'équation 3.5 est utilisée car elle respecte cette propriété en plus de diminuer l'étalement spectral [45]. La forme de cette fenêtre est illustrée à la figure 3.3.

$$(w[n])^2 + (w[\text{mod}\{n + (N)(o), N\}])^2 = 1 \quad 0 \leq n < N \quad (3.4)$$

$$w[n] = \begin{cases} 0.5 - 0.5 \cos(7.9916n/N) & 0 \leq n < N/4 \\ \sqrt{1 - (0.5 - 0.5 \cos(3.9958 - (7.9916n/N)))^2} & N/4 \leq n < N/2 \\ \sqrt{1 - (0.5 - 0.5 \cos((7.9916n/N) - 3.9958))^2} & N/2 \leq n < 3N/4 \\ 0.5 - 0.5 \cos(7.9916 - 7.9916n/N) & 3N/4 \leq n < N \end{cases} \quad (3.5)$$

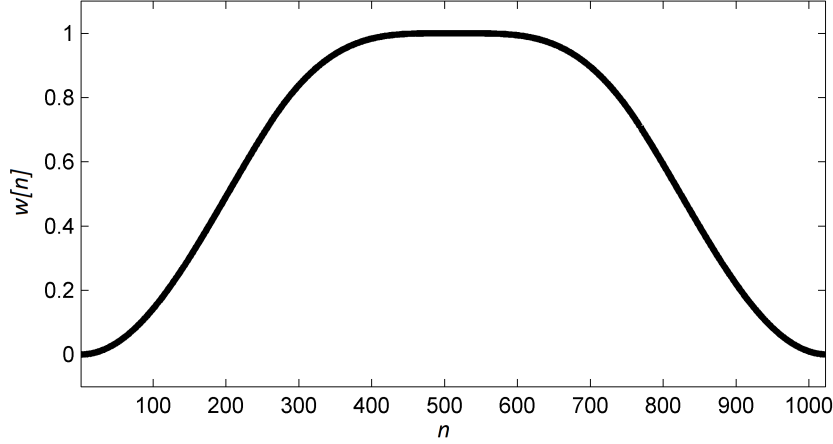


Figure 3.3 Fenêtre d'analyse et de synthèse

Pour chaque trame, la transformée de Fourier discrète est calculée tel qu'exprimée par l'équation 3.6. La variable k représente l'indice de chaque fréquence discrétisée.

$$(X_{mic})_m^l[k] = \sum_{n=0}^{N-1} (x_{mic})_m^l[n] e^{-j2\pi nk} \quad 0 \leq k < N \quad (3.6)$$

3.1.2 Pondération et corrélation croisée

La corrélation croisée $R_{m_1, m_2}^l(\tau)$ entre les signaux des microphones m_1 et m_2 pour la trame l est obtenue grâce à l'équation 3.7. Les opérateurs $|(.)|$ et $(.)^*$ représentent la magnitude et la conjuguée complexe respectivement. La variable τ est un nombre entier qui représente le délai en échantillons. Le spectre est normalisé afin d'augmenter la résolution spatiale. En effet, la majeure partie de l'énergie de la voix se situe dans les fréquences relativement basses, ce qui entraîne une faible résolution spatiale [58]. Cependant, cette normalisation augmente la contribution du bruit dans certaines régions du spectre. Pour y remédier, le spectre de chaque microphone est multiplié par un facteur de pondération $\zeta_m^l[k]$.

$$R_{m_1, m_2}^l(\tau) = \sum_{k=0}^{N-1} \left(\frac{\zeta_{m_1}^l[k] (X_{mic})_{m_1}^l[k]}{|(X_{mic})_{m_1}^l[k]|} \right) \left(\frac{\zeta_{m_2}^l[k] (X_{mic})_{m_2}^l[k]}{|(X_{mic})_{m_2}^l[k]|} \right)^* e^{(j2\pi k\tau/N)} \quad (3.7)$$

Le facteur de pondération $\zeta_m^l[k]$ est calculé selon l'équation 3.8 à partir de l'estimation du rapport signal sur bruit *a priori* $(\xi_{mic})_m^l[k]$ [12] obtenu grâce à l'équation 3.9. La constante α_d et la variable $(\lambda_{mic})_m^l[k]$ représentent le taux d'adaptation et le bruit, respectivement. Pour obtenir des résultats optimaux, il a été déterminé expérimentalement que $\alpha_d = 0.95$. Les variables m , l et k représentent les indices du microphone, de la trame et de la fréquence discrétisée respectivement.

$$\zeta_m^l[k] = \frac{(\xi_{mic})_m^l[k]}{(\xi_{mic})_m^l[k] + 1} \quad (3.8)$$

$$(\xi_{mic})_m^l[k] = \frac{(1 - \alpha_d)(\zeta_m^{l-1}[k])^2 |(X_{mic})_m^{l-1}[k]|^2 + \alpha_d |(X_{mic})_m^l[k]|^2}{(\lambda_{mic})_m^l[k]} \quad (3.9)$$

Le bruit total est le résultat de l'addition de la réverbération $(\lambda_{mic}^{rev})_m^l[k]$ et du bruit stationnaire de l'environnement $(\lambda_{mic}^{stat})_m^l[k]$ tel qu'exprimé en 3.10.

$$(\lambda_{mic})_m^l[k] = (\lambda_{mic}^{rev})_m^l[k] + (\lambda_{mic}^{stat})_m^l[k] \quad (3.10)$$

Les premières réflexions sont ignorées tandis que la réverbération tardive est estimée à l'aide de l'équation 3.11. Le temps de réverbération de la pièce (T_{60}) est représenté par la constante γ et le rapport signal sur réverbération correspond à la constante δ . Pour obtenir des résultats optimaux, il a été déterminé expérimentalement que $\gamma = 0.3$ et $\delta = 1.0$.

$$(\lambda_{mic}^{rev})_m^l[k] = \gamma (\lambda_{mic}^{rev})_m^{l-1}[k] + \left(\frac{(1 - \gamma)}{\delta} \right) |\zeta_m^{l-1}[k]| (X_{mic})_m^{l-1}[k]|^2 \quad (3.11)$$

Pour estimer le bruit stationnaire, la méthode de moyenne récursive obtenue à l'aide des minima (*Minima-Controlled Recursive Average* (MCRA)) est employée [8]. Le signal est d'abord filtré dans le domaine du temps, tel que montré en 3.12. L'opérateur $(*)$ symbolise la convolution linéaire. Puisque $(X_{mic})_m^l[k]$ est seulement défini dans l'intervalle $0 \leq k < N$, cette variable se voit attribuée une valeur de zéro pour $k < 0$ et $k \geq N$. La variable $[(\lambda_{mic}^{stat})_t]_m^l[k]$ représente le spectre filtré dans le domaine temporel. La fenêtre $b[k]$ utilisée est présentée par l'équation 3.13.

$$[(\lambda_{mic}^{stat})_t]_m^l[k] = b[k] * |(X_{mic})_m^l[k]|^2 \quad (3.12)$$

$$b[k] = \begin{cases} 0.25 & k = -1 \\ 0.50 & k = 0 \\ 0.25 & k = 1 \\ 0 & \text{autrement} \end{cases} \quad (3.13)$$

Le signal est par la suite filtré dans le domaine fréquentiel en 3.14. La constante α_s représente le taux d'adaptation. Elle fut fixée expérimentalement à $\alpha_s = 0.95$.

$$[(\lambda_{mic}^{stat})_f]_m^l[k] = \alpha_s [(\lambda_{mic}^{stat})_f]_m^l[k] + (1 - \alpha_s) [(\lambda_{mic}^{stat})_t]_m^l[k] \quad (3.14)$$

Les variables $[(\lambda_{mic}^{stat})_{min}]_m^l[k]$ et $[(\lambda_{mic}^{stat})_{tmp}]_m^l[k]$ sont ensuite mises à jour telles que l'expriment les équations 3.15 et 3.16. La variable L_{mcra} réfère au nombre de trames utilisées pour évaluer le bruit stationnaire. Dans le cas présent, $L_{mcra} = 150$. Les variables m , l et k représentent les indices du microphone, de la trame et de la fréquence discrétisée respectivement.

$$[(\lambda_{mic}^{stat})_{min}]_m^l[k] = \begin{cases} \min\{[(\lambda_{mic}^{stat})_{tmp}]_m^l[k], [(\lambda_{mic}^{stat})_f]_m^l[k]\} & (\text{mod}(l, L_{mcra}) = 0) \\ \min\{[(\lambda_{mic}^{stat})_{min}]_m^l[k], [(\lambda_{mic}^{stat})_f]_m^l[k]\} & (\text{mod}(l, L_{mcra}) \neq 0) \end{cases} \quad (3.15)$$

$$[(\lambda_{mic}^{stat})_{tmp}]_m^l[k] = \begin{cases} [(\lambda_{mic}^{stat})_f]_m^l[k] & (\text{mod}(l, L_{mcra}) = 0) \\ \min\{[(\lambda_{mic}^{stat})_{tmp}]_m^l[k], [(\lambda_{mic}^{stat})_f]_m^l[k]\} & (\text{mod}(l, L_{mcra}) \neq 0) \end{cases} \quad (3.16)$$

Le facteur de mises à jour du bruit α_u est ensuite défini dans l'équation 3.17, avec $\delta = 1.0$ déterminé expérimentalement.

$$\alpha_u = \begin{cases} \max\{(1/l), \alpha_d\} & \left\{ \begin{array}{ll} l < L_{mcra} & \text{ou} \\ [(\lambda_{mic}^{stat})_f]_m^l[k] < (\delta [(\lambda_{mic}^{stat})_{min}]_m^{l-1}[k]) & \text{ou} \\ (\lambda_{mic}^{stat})_m^{l-1}[k] > |(X_{mic})_m^l[k]|^2 & \end{array} \right. \\ 1 & \text{autrement} \end{cases} \quad (3.17)$$

Une fois le facteur déterminé, le bruit $(\lambda_{mic}^{stat})_m^l[k]$ est estimé grâce à l'équation 3.18. Les variables m , l et k représentent les indices du microphone, de la trame et de la fréquence discrétisée respectivement.

$$(\lambda_{mic}^{stat})_m^l[k] = \alpha_u (\lambda_{mic}^{stat})_m^{l-1}[k] + (1 - \alpha_u) |(X_{mic})_m^l[k]|^2 \quad (3.18)$$

3.1.3 Sphère unitaire

L'espace de recherche pour les sources sonores est discrétisée à l'aide d'une sphère unitaire, dont chaque point représente une position potentielle pour une source [45]. Cette sphère est constituée de triangles $(\Delta_{\text{sph}})_v^{level}$. La variable v représente l'indice de chaque triangle, pour un total de Υ_{level} triangles à chaque niveau $level$. Les variables $[(\Delta_{\text{sph}})_v^{level}]_1$, $[(\Delta_{\text{sph}})_v^{level}]_2$ et $[(\Delta_{\text{sph}})_v^{level}]_3$ représentent chacun des points qui forment les sommets du triangle, tel que l'exprime l'équation 3.19.

$$(\Delta_{\text{sph}})_v^{level} = \begin{bmatrix} [(\Delta_{\text{sph}})_v^{level}]_1 \\ [(\Delta_{\text{sph}})_v^{level}]_2 \\ [(\Delta_{\text{sph}})_v^{level}]_3 \end{bmatrix} \quad (3.19)$$

Pour chaque ensemble de Υ_{level} triangles, il est possible d'extraire $D_{level} = (\Upsilon_{level}/2 + 2)$ points puisque les triangles partagent plusieurs points. Ces points $(\mathbf{z}_{\text{sph}})_d^{level}$ sont répartis de manière uniforme, à un niveau $level$ et un indice d pour un total de D_{level} points. Les variables $[(z_{sph})_d^{level}]_x$, $[(z_{sph})_d^{level}]_y$ et $[(z_{sph})_d^{level}]_z$ représentent les coordonnées en x , y et z du point $(\mathbf{z}_{\text{sph}})_d^{level}$ respectivement, selon l'équation 3.20.

$$(\mathbf{z}_{\text{sph}})_d^{level} = \begin{bmatrix} [(z_{sph})_d^{level}]_x \\ [(z_{sph})_d^{level}]_y \\ [(z_{sph})_d^{level}]_z \end{bmatrix} \quad (3.20)$$

Une sphère initiale icosaédrale est générée au niveau $level = 0$ avec $D_0 = 12$ et $\Upsilon_0 = 20$. Les points et les triangles initiaux sont disponibles dans les tableaux 3.1 et 3.2. Pour chaque niveau, quatre triangles sont générés à partir du triangle initial. Les points du triangle initial sont définis par les équations 3.21, 3.22 et 3.23.

$$[(\Delta_{\mathbf{sph}})_v^{level}]_1 = \begin{bmatrix} \left([(\Delta_{sph})_v^{level}]_1 \right)_x \\ \left([(\Delta_{sph})_v^{level}]_1 \right)_y \\ \left([(\Delta_{sph})_v^{level}]_1 \right)_z \end{bmatrix} \quad (3.21)$$

$$[(\Delta_{\mathbf{sph}})_v^{level}]_2 = \begin{bmatrix} \left([(\Delta_{sph})_v^{level}]_2 \right)_x \\ \left([(\Delta_{sph})_v^{level}]_2 \right)_y \\ \left([(\Delta_{sph})_v^{level}]_2 \right)_z \end{bmatrix} \quad (3.22)$$

$$[(\Delta_{\mathbf{sph}})_v^{level}]_3 = \begin{bmatrix} \left([(\Delta_{sph})_v^{level}]_3 \right)_x \\ \left([(\Delta_{sph})_v^{level}]_3 \right)_y \\ \left([(\Delta_{sph})_v^{level}]_3 \right)_z \end{bmatrix} \quad (3.23)$$

Tableau 3.1 Points initiaux pour la sphère icosaédrale

d	$[(z_{sph})_d^0]_x$	$[(z_{sph})_d^0]_y$	$[(z_{sph})_d^0]_z$
0	+0.0000	+0.0000	+1.0000
1	+0.0000	+0.8944	+0.4472
2	+0.8507	+0.2764	+0.4472
3	+0.5257	-0.7236	+0.4472
4	-0.5257	-0.7236	+0.4472
5	-0.8507	+0.2764	+0.4472
6	+0.0000	+0.0000	-1.0000
7	+0.0000	-0.8944	-0.4472
8	-0.8507	-0.2764	-0.4472
9	-0.5257	+0.7236	-0.4472
10	+0.5257	+0.7236	-0.4472
11	+0.8507	-0.2764	-0.4472

Par la suite, chaque paire de points est additionnée et normalisée pour former des points intermédiaires tels que démontrés dans les équations 3.24, 3.25 et 3.26. L'opérateur $\|.\|$ retourne la magnitude du vecteur.

$$[(\Delta_{\mathbf{sph}})_v^{level}]_A = \begin{bmatrix} \frac{\left([(\Delta_{sph})_v^{level}]_1 \right)_x + \left([(\Delta_{sph})_v^{level}]_2 \right)_x}{\|[(\Delta_{\mathbf{sph}})_v^{level}]_1 + [(\Delta_{\mathbf{sph}})_v^{level}]_2\|} \\ \frac{\left([(\Delta_{sph})_v^{level}]_1 \right)_y + \left([(\Delta_{sph})_v^{level}]_2 \right)_y}{\|[(\Delta_{\mathbf{sph}})_v^{level}]_1 + [(\Delta_{\mathbf{sph}})_v^{level}]_2\|} \\ \frac{\left([(\Delta_{sph})_v^{level}]_1 \right)_z + \left([(\Delta_{sph})_v^{level}]_2 \right)_z}{\|[(\Delta_{\mathbf{sph}})_v^{level}]_1 + [(\Delta_{\mathbf{sph}})_v^{level}]_2\|} \end{bmatrix} \quad (3.24)$$

$$[(\Delta_{\text{sph}})_v^{level}]_{\text{B}} = \begin{bmatrix} \frac{([\Delta_{sph})_v^{level}]_2)_x + ([\Delta_{sph})_v^{level}]_3)_x}{\|[(\Delta_{\text{sph}})_v^{level}]_2 + [(\Delta_{\text{sph}})_v^{level}]_3\|} \\ \frac{([\Delta_{sph})_v^{level}]_2)_y + ([\Delta_{sph})_v^{level}]_3)_y}{\|[(\Delta_{\text{sph}})_v^{level}]_2 + [(\Delta_{\text{sph}})_v^{level}]_3\|} \\ \frac{([\Delta_{sph})_v^{level}]_2)_z + ([\Delta_{sph})_v^{level}]_3)_z}{\|[(\Delta_{\text{sph}})_v^{level}]_2 + [(\Delta_{\text{sph}})_v^{level}]_3\|} \end{bmatrix} \quad (3.25)$$

$$[(\Delta_{\text{sph}})_v^{level}]_{\text{C}} = \begin{bmatrix} \frac{([\Delta_{sph})_v^{level}]_3)_x + ([\Delta_{sph})_v^{level}]_1)_x}{\|[(\Delta_{\text{sph}})_v^{level}]_3 + [(\Delta_{\text{sph}})_v^{level}]_1\|} \\ \frac{([\Delta_{sph})_v^{level}]_3)_y + ([\Delta_{sph})_v^{level}]_1)_y}{\|[(\Delta_{\text{sph}})_v^{level}]_3 + [(\Delta_{\text{sph}})_v^{level}]_1\|} \\ \frac{([\Delta_{sph})_v^{level}]_3)_z + ([\Delta_{sph})_v^{level}]_1)_z}{\|[(\Delta_{\text{sph}})_v^{level}]_3 + [(\Delta_{\text{sph}})_v^{level}]_1\|} \end{bmatrix} \quad (3.26)$$

Tableau 3.2 Triangles initiaux pour la sphère icosaédrale

v	$[(\Delta_{\text{sph}})_v^{level}]_1$	$[(\Delta_{\text{sph}})_v^{level}]_2$	$[(\Delta_{\text{sph}})_v^{level}]_3$
0	$(\mathbf{Z}_{\text{sph}})_0$	$(\mathbf{Z}_{\text{sph}})_1$	$(\mathbf{Z}_{\text{sph}})_2$
1	$(\mathbf{Z}_{\text{sph}})_0$	$(\mathbf{Z}_{\text{sph}})_2$	$(\mathbf{Z}_{\text{sph}})_3$
2	$(\mathbf{Z}_{\text{sph}})_0$	$(\mathbf{Z}_{\text{sph}})_3$	$(\mathbf{Z}_{\text{sph}})_4$
3	$(\mathbf{Z}_{\text{sph}})_0$	$(\mathbf{Z}_{\text{sph}})_4$	$(\mathbf{Z}_{\text{sph}})_5$
4	$(\mathbf{Z}_{\text{sph}})_0$	$(\mathbf{Z}_{\text{sph}})_5$	$(\mathbf{Z}_{\text{sph}})_1$
5	$(\mathbf{Z}_{\text{sph}})_1$	$(\mathbf{Z}_{\text{sph}})_{10}$	$(\mathbf{Z}_{\text{sph}})_2$
6	$(\mathbf{Z}_{\text{sph}})_2$	$(\mathbf{Z}_{\text{sph}})_{10}$	$(\mathbf{Z}_{\text{sph}})_{11}$
7	$(\mathbf{Z}_{\text{sph}})_2$	$(\mathbf{Z}_{\text{sph}})_{11}$	$(\mathbf{Z}_{\text{sph}})_3$
8	$(\mathbf{Z}_{\text{sph}})_3$	$(\mathbf{Z}_{\text{sph}})_{11}$	$(\mathbf{Z}_{\text{sph}})_7$
9	$(\mathbf{Z}_{\text{sph}})_3$	$(\mathbf{Z}_{\text{sph}})_7$	$(\mathbf{Z}_{\text{sph}})_4$
10	$(\mathbf{Z}_{\text{sph}})_4$	$(\mathbf{Z}_{\text{sph}})_7$	$(\mathbf{Z}_{\text{sph}})_8$
11	$(\mathbf{Z}_{\text{sph}})_4$	$(\mathbf{Z}_{\text{sph}})_8$	$(\mathbf{Z}_{\text{sph}})_5$
12	$(\mathbf{Z}_{\text{sph}})_5$	$(\mathbf{Z}_{\text{sph}})_8$	$(\mathbf{Z}_{\text{sph}})_9$
13	$(\mathbf{Z}_{\text{sph}})_5$	$(\mathbf{Z}_{\text{sph}})_9$	$(\mathbf{Z}_{\text{sph}})_1$
14	$(\mathbf{Z}_{\text{sph}})_1$	$(\mathbf{Z}_{\text{sph}})_9$	$(\mathbf{Z}_{\text{sph}})_{10}$
15	$(\mathbf{Z}_{\text{sph}})_6$	$(\mathbf{Z}_{\text{sph}})_{11}$	$(\mathbf{Z}_{\text{sph}})_{10}$
16	$(\mathbf{Z}_{\text{sph}})_6$	$(\mathbf{Z}_{\text{sph}})_7$	$(\mathbf{Z}_{\text{sph}})_{11}$
17	$(\mathbf{Z}_{\text{sph}})_6$	$(\mathbf{Z}_{\text{sph}})_8$	$(\mathbf{Z}_{\text{sph}})_7$
18	$(\mathbf{Z}_{\text{sph}})_6$	$(\mathbf{Z}_{\text{sph}})_9$	$(\mathbf{Z}_{\text{sph}})_8$
19	$(\mathbf{Z}_{\text{sph}})_6$	$(\mathbf{Z}_{\text{sph}})_{10}$	$(\mathbf{Z}_{\text{sph}})_9$

Les nouveaux points obtenus sont alors utilisés pour former de nouveaux triangles définis par les équations 3.27 à 3.30.

$$\begin{aligned} [(\Delta_{\text{sph}})_{4v+0}^{level+1}]_1 &= [(\Delta_{\text{sph}})_v^{level}]_1 \\ [(\Delta_{\text{sph}})_{4v+0}^{level+1}]_2 &= [(\Delta_{\text{sph}})_v^{level}]_A \\ [(\Delta_{\text{sph}})_{4v+0}^{level+1}]_3 &= [(\Delta_{\text{sph}})_v^{level}]_C \end{aligned} \quad (3.27)$$

$$\begin{aligned} [(\Delta_{\text{sph}})_{4v+1}^{level+1}]_1 &= [(\Delta_{\text{sph}})_v^{level}]_C \\ [(\Delta_{\text{sph}})_{4v+1}^{level+1}]_2 &= [(\Delta_{\text{sph}})_v^{level}]_B \\ [(\Delta_{\text{sph}})_{4v+1}^{level+1}]_3 &= [(\Delta_{\text{sph}})_v^{level}]_3 \end{aligned} \quad (3.28)$$

$$\begin{aligned} [(\Delta_{\text{sph}})_{4v+2}^{level+1}]_1 &= [(\Delta_{\text{sph}})_v^{level}]_C \\ [(\Delta_{\text{sph}})_{4v+2}^{level+1}]_2 &= [(\Delta_{\text{sph}})_v^{level}]_A \\ [(\Delta_{\text{sph}})_{4v+2}^{level+1}]_3 &= [(\Delta_{\text{sph}})_v^{level}]_B \end{aligned} \quad (3.29)$$

$$\begin{aligned} [(\Delta_{\text{sph}})_{4v+3}^{level+1}]_1 &= [(\Delta_{\text{sph}})_v^{level}]_A \\ [(\Delta_{\text{sph}})_{4v+3}^{level+1}]_2 &= [(\Delta_{\text{sph}})_v^{level}]_2 \\ [(\Delta_{\text{sph}})_{4v+3}^{level+1}]_3 &= [(\Delta_{\text{sph}})_v^{level}]_B \end{aligned} \quad (3.30)$$

Ce processus itératif est répété jusqu'au niveau 3 ($level = 3$). Il existe alors $\Upsilon_3 = 5120$ triangles, ce qui permet d'extraire $D_3 = 2562$ points répartis uniformément sur une sphère unitaire. Pour chaque point d sur la sphère, le délai d'arrivée $delay_{m_1, m_2}(d)$ du signal entre les microphones m_1 et m_2 est calculé à l'aide de l'équation 3.31.

$$delay_{m_1, m_2}(d) = \text{round} \left[\left(\frac{F_s}{c_{air}} \right) (\|(\mathbf{z}_{\text{sph}})_d^3 - (\mathbf{z}_{\text{mic}})_{m_2}\| - \|(\mathbf{z}_{\text{sph}})_d^3 - (\mathbf{z}_{\text{mic}})_{m_1}\|) \right] \quad (3.31)$$

L'opérateur $\| \cdot \|$ retourne la magnitude du vecteur. Les variable F_s et c_{air} représentent le taux d'échantillonnage et la vitesse du son dans l'air respectivement. Les valeurs de $delay_{m_1, m_2}(d)$ sont calculées une seule fois au début de l'exécution et mises en mémoire.

3.1.4 Recherche

Les positions potentielles des sources sonores sont évaluées grâce à l'algorithme suivant [45]. Un nombre fixe de sources sonores ($Q = 4$) est utilisé même si ce nombre ne correspond pas forcément au nombre réel de sources présentes. Il fut démontré expérimentalement qu'une valeur de Q inférieure à 4 diminue les performances du système lorsque plusieurs sources sont actives simultanément, tandis qu'une valeur supérieure à 4 augmente la charge de cal-

culs sans améliorer significativement la localisation [45]. Le filtre particulaire présenté à la section 3.2 règle ce problème en gérant les fausses détections. La variable $E_q^l(d)$ représente l'énergie du formateur de faisceaux à la position $(\mathbf{z}_{\text{sph}})_d^3$ pour l'itération q de la trame l . Les variables $(\mathbf{z}_{\text{pot}})_q^l$ et $(E_{\text{pot}})_q^l$ représentent la position et l'énergie de chaque source potentielle q respectivement. Après chaque itération, certaines valeurs de corrélation sont remises à zéro dans le but de réduire la contribution énergétique de la source potentielle à cette position pour les prochaines itérations. La variable τ_{range} représente l'étendue de la plage de cette remise à zéro. Dans le cas présent, $\tau_{\text{range}} = 5$.

Algorithme de localisation de sources multiples

Calcul de $R_{m_1, m_2}^l(\tau)$ pour toutes les valeurs de τ , m_1 et m_2 .

Pour $q = 0$ **jusqu'à** $q = Q - 1$

Pour $d = 0$ **jusqu'à** $d = D_3 - 1$

$$E_q^l(d) = 0$$

Pour $m_1 = 0$ **jusqu'à** $(M - 2)$

Pour $m_2 = (m_1 + 1)$ **jusqu'à** $(M - 1)$

$$E_q^l(d) = E_q^l(d) + R_{m_1, m_2}^l(\text{delay}_{m_1, m_2}(d))$$

Fin de la boucle

Fin de la boucle

Fin de la boucle

$$(d_{\text{max}})_q^l = \arg \max_d (E_q^l(d))$$

$$(\mathbf{z}_{\text{pot}})_q^l = (\mathbf{z}_{\text{sph}})_{d_{\text{max}}}^3$$

$$(E_{\text{pot}})_q^l = E_q^l(d_{\text{max}})$$

Pour $m_1 = 0$ **jusqu'à** $(M - 2)$

Pour $m_2 = (m_1 + 1)$ **jusqu'à** $(M - 1)$

$$(\tau_{\text{max}})_q^l = \text{delay}_{m_1, m_2}((d_{\text{max}})_q^l)$$

Pour $\tau' = ((\tau_{\text{max}})_q^l - \tau_{\text{range}})$ **jusqu'à** $\tau' = ((\tau_{\text{max}})_q^l + \tau_{\text{range}})$

$$R_{m_1, m_2}^l(\tau') = 0$$

Fin de la boucle

Fin de la boucle

Fin de la boucle

Fin de la boucle

3.2 Suivi de sources sonores avec ManyEars

Le suivi, la détection et l'élimination de sources sonores sont effectués à l'aide de filtres particulières [49]. Chaque source suivie est modélisée par un filtre qui est un composé de $F = 500$ particules. Chaque particule possède une position, une vitesse et une pondération. La moyenne pondérée des positions de toutes les particules d'un filtre représente une estimation de la position actuelle de la source sonore. Ces filtres particuliers sont mis à jour à chaque nouvelle trame pour tenir compte des nouvelles observations fournies par le formateur de faisceaux.

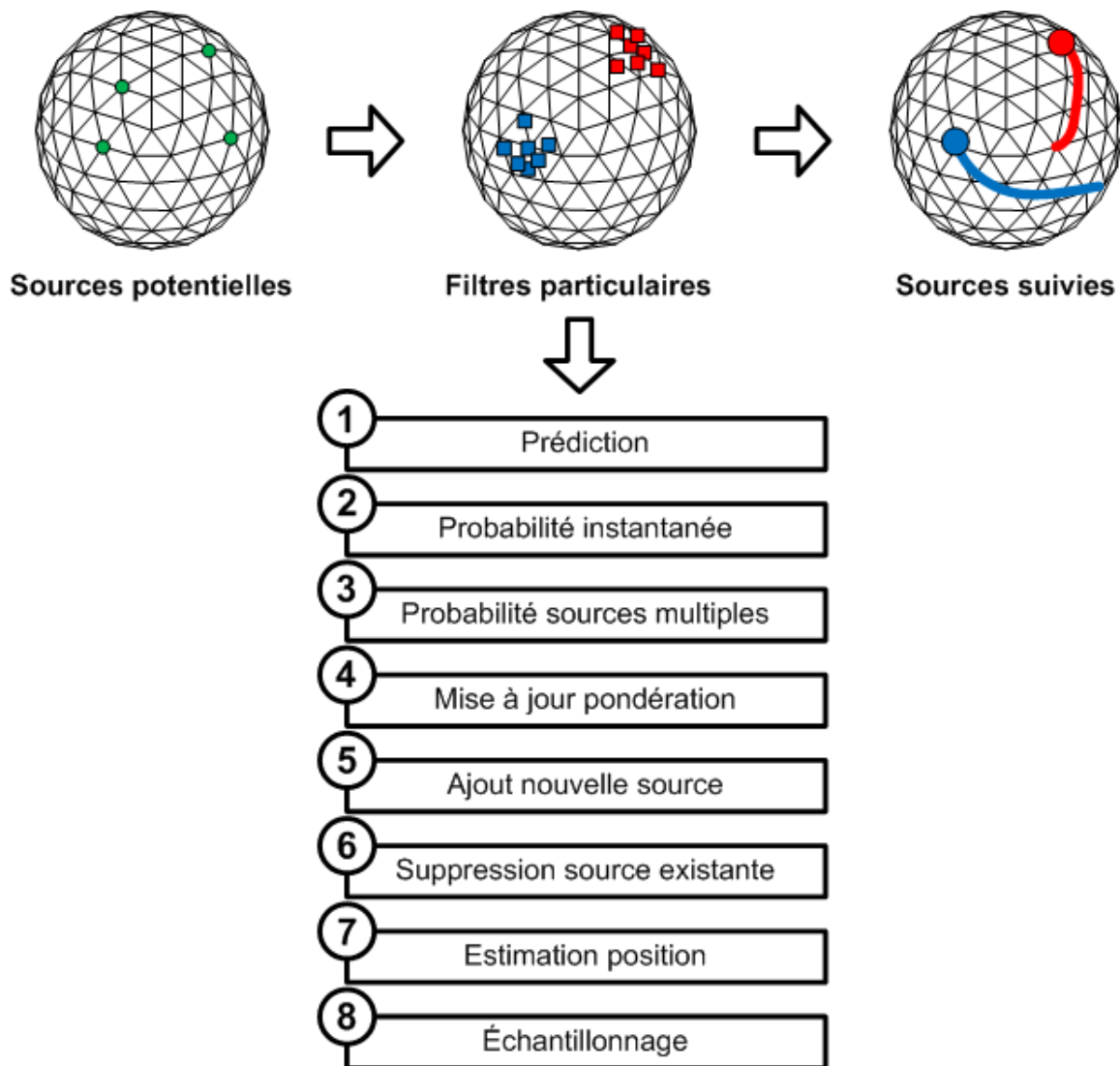


Figure 3.4 Aperçu du mécanisme de suivi de ManyEars

3.2.1 Prédiction

Chaque particule f au sein d'un filtre s appliqué à une trame l possède une position $(\mathbf{z}_{\text{part}})_s^l(f)$ et une vitesse $(\mathbf{z}_{\text{part}})_s^l(f)$ telles qu'exprimées dans les équations 3.32 et 3.33.

$$(\mathbf{z}_{\text{part}})_s^l(f) = \begin{bmatrix} [(z_{\text{part}})_s^l(f)]_x \\ [(z_{\text{part}})_s^l(f)]_y \\ [(z_{\text{part}})_s^l(f)]_z \end{bmatrix} \quad 0 \leq f < F \quad (3.32)$$

$$(\mathbf{z}_{\text{part}})_s^l(f) = \begin{bmatrix} [(\dot{z}_{\text{part}})_s^l(f)]_x \\ [(\dot{z}_{\text{part}})_s^l(f)]_y \\ [(\dot{z}_{\text{part}})_s^l(f)]_z \end{bmatrix} \quad 0 \leq f < F \quad (3.33)$$

Les positions et les vitesses sont mises à jour récursivement selon une excitation stochastique donnée en 3.34 et 3.35. La constante ΔT représente l'intervalle de temps entre les mises à jour (la notion de temps est utilisée par rapport à la vitesse de la particule, et non forcément par rapport au temps réel qui s'est écoulé entre deux mises à jour). Dans le cas présent, $\Delta T = 0.008$. La variable F_x correspond à la réalisation d'une variable aléatoire avec une distribution normale.

$$(\mathbf{z}_{\text{part}'})_s^l(f) = (\mathbf{z}_{\text{part}})_s^{l-1}(f) + \Delta T (\mathbf{z}_{\text{part}'})_s^l(f) \quad (3.34)$$

$$(\mathbf{z}_{\text{part}'})_s^l(f) = (a_s^l(f))(\mathbf{z}_{\text{part}})_s^{l-1}(f) + (b_s^l(f))F_x \quad (3.35)$$

La position et la vitesse de chaque particule sont normalisées de manière à ce que chaque particule soit située sur la sphère unitaire et que sa vitesse soit tangentielle. Ces opérations de normalisation sont décrites dans les équations 3.36 et 3.37. L'opérateur (\cdot) symbolise le produit scalaire de deux vecteurs.

$$(\mathbf{z}_{\text{part}})_s^l(f) = \frac{(\mathbf{z}_{\text{part}'})_s^l(f)}{\|(\mathbf{z}_{\text{part}'})_s^l(f)\|} \quad (3.36)$$

$$(\mathbf{z}_{\text{part}})_s^l(f) = (\mathbf{z}_{\text{part}'})_s^l(f) - [(\mathbf{z}_{\text{part}})_s^l(f) \cdot (\mathbf{z}_{\text{part}'})_s^l(f)](\mathbf{z}_{\text{part}})_s^l(f) \quad (3.37)$$

Les paramètres $a_s^l(f)$ et $b_s^l(f)$ sont définis dans les équations 3.38 et 3.39 respectivement. Les paramètres $\alpha_s^l(f)$ et $\beta_s^l(f)$ sont définis en fonction de l'état de chaque particule. Les variables s , l et f représentent les indices du filtre, de la trame et de la particule respec-

tivement. Une particule peut être en arrêt, à vitesse constante ou en accélération. L'état de chaque particule est déterminé par la réalisation d'un processus aléatoire $\kappa_s^l(f)$ blanc avec une distribution uniforme entre 0 et 1. Les probabilités pour un arrêt, une vitesse constante ou une accélération sont définies par les constantes p_{stp} , p_{cst} et p_{acc} respectivement. Évidemment, la somme de ces trois constantes équivaut à une valeur de 1. Dans le cas présent, $p_{stp} = 0.50$, $p_{cst} = 0.20$ et $p_{acc} = 0.30$. Le tableau 3.3 illustre l'attribution des paramètres.

Tableau 3.3 Paramètres des particules

$\kappa_s^l(f)$	État	$\alpha_s^l(f)$	$\beta_s^l(f)$
$[0, p_{acc}]$	Arrêt	2	0.04
$[p_{acc}, (p_{acc} + p_{cst})]$	Vitesse constante	0.05	0.2
$[(p_{acc} + p_{cst}), 1]$	Accélération	0.5	0.2

$$a_s^l(f) = e^{-\alpha_s^l(f)\Delta T} \quad (3.38)$$

$$b_s^l(f) = \beta_s^l(f)\sqrt{1 - a_s^l(f)^2} \quad (3.39)$$

3.2.2 Probabilité instantanée

La densité de probabilité d'observer une source potentielle O_q^l à la position d'une particule existante $(\mathbf{z}_{\text{part}})_s^l(f)$ a été déterminée expérimentalement et correspond à une somme de distributions normales telle que démontrée dans les équations 3.40 et 3.41.

$$dist = \|(\mathbf{z}_{\text{part}})_s^l(f) - (\mathbf{z}_{\text{pot}})_q^l\| \quad (3.40)$$

$$p(O_q^l | (\mathbf{z}_{\text{part}})_s^l(f)) = 0.8e^{-80dist} + 0.18e^{-8dist} + 0.02e^{-0.4dist} \quad (3.41)$$

3.2.3 Probabilité pour des sources multiples

Pour chaque source potentielle observée à partir du formateur de faisceaux, trois hypothèses sont avancées :

- H_0 : Il s'agit d'une fausse détection.
- H_1 : Il s'agit d'une source déjà suivie.
- H_2 : Il s'agit d'une nouvelle source.

Une variable $g_c^l(q)$ est alors utilisée pour relier les sources potentielles à des sources suivies. Les valeurs -2 et -1 représentent une fausse détection et une nouvelle source respectivement. Les valeurs 0 à $(S - 1)$ représentent l'indice de la source suivie. Cette relation est montrée par l'équation 3.42.

$$g_c^l(q) \in \{-2, -1, 0, 1, \dots, S - 1\} \quad (3.42)$$

Le vecteur \mathbf{g}_c^l représente la relation c qui fait partie d'un ensemble G^l de C réalisations valides pour l'attribution des hypothèses pour toutes les sources potentielles, telles que présentées par les équations 3.43 et 3.44. Les variables c , l et q représentent donc les indices de la réalisation, la trame et la source potentielle respectivement.

$$\mathbf{g}_c^l = \begin{bmatrix} g_c^l(0) \\ g_c^l(1) \\ \vdots \\ g_c^l(Q - 1) \end{bmatrix} \quad (3.43)$$

$$G^l = \{\mathbf{g}_0^l, \mathbf{g}_1^l, \dots, \mathbf{g}_{C-1}^l\} \quad (3.44)$$

Pour établir la valeur de C , il s'agit de trouver Q^{S+2} possibilités et de soustraire les cas impossibles lorsque plusieurs sources potentielles sont assignées à la même source suivie. Un exemple est fourni dans le tableau 3.4 avec $Q = 2$ et $S = 2$. Une fois les combinaisons impossibles retirées, l'ensemble G^l est composé de $C = 14$ vecteurs tels qu'illustrés dans l'équation 3.45.

Tableau 3.4 Exemple pour l'ensemble G^l

Valide	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	
$\mathbf{g}_c^l(0)$	-2	-1	0	1	-2	-1	0	1	-2	-1	0	1	-2	-1	0	1
$\mathbf{g}_c^l(1)$	-2	-2	-2	-2	-1	-1	-1	-1	0	0	0	0	1	1	1	1

$$G^l = \left\{ \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \begin{bmatrix} -1 \\ -2 \end{bmatrix}, \begin{bmatrix} 0 \\ -2 \end{bmatrix}, \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \begin{bmatrix} -2 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\} \quad (3.45)$$

Une fois l'ensemble G^l défini, la probabilité non normalisée que la source potentielle q soit une fausse détection est définie dans l'équation 3.46. La variable $P(\mathbf{g}_c^l|O^l)$ représente la probabilité d'obtenir la réalisation de \mathbf{g}_c^l si l'observation O^l est connue. Il s'agit donc de calculer la somme de $P(\mathbf{g}_c^l|O^l)$ pour tous les vecteurs \mathbf{g}_c^l de l'ensemble G^l qui contiennent une assignation à une fausse détection pour la source potentielle q . De la même manière, la probabilité non normalisée que la source potentielle q soit déjà suivie par le filtre s est calculée dans l'équation 3.47. Finalement, la probabilité non normalisée que la source potentielle soit une nouvelle source qui n'était pas suivie jusqu'à présent est donnée dans l'équation 3.48.

$$(P_{H_0}^l)'(q) = \sum_{c=0}^{C-1} \delta_{-2, g_c^l(q)} P(\mathbf{g}_c^l|O^l) \quad (3.46)$$

$$(P_s^l)'(q) = \sum_{c=0}^{C-1} \delta_{s, g_c^l(q)} P(\mathbf{g}_c^l|O^l) \quad (3.47)$$

$$(P_{H_2}^l)'(q) = \sum_{c=0}^{C-1} \delta_{-1, g_c^l(q)} P(\mathbf{g}_c^l|O^l) \quad (3.48)$$

La fonction $\delta_{x,y}$ est définie par l'équation 3.49.

$$\delta_{x,y} = \begin{cases} 0 & x \neq y \\ 1 & x = y \end{cases} \quad (3.49)$$

Ces probabilités sont ensuite normalisées par les équations 3.50 à 3.53.

$$P_{tot} = \sum_{q=0}^{Q-1} \left[(P_{H_0}^l)'(q) + (P_{H_2}^l)'(q) + \sum_{s=0}^{S-1} (P_s^l)'(q) \right] \quad (3.50)$$

$$P_{H_0}^l(q) = \frac{(P_{H_0}^l)'(q)}{P_{tot}} \quad (3.51)$$

$$P_s^l(q) = \frac{(P_s^l)'(q)}{P_{tot}} \quad (3.52)$$

$$P_{H_2}^l(q) = \frac{(P_{H_2}^l)'(q)}{P_{tot}} \quad (3.53)$$

Par la suite, puisque chaque assignation est indépendante, $P(\mathbf{g}_c^l|O^l)$ peut être calculée à partir de l'équation 3.54.

$$P(\mathbf{g}_c^l|O^l) = p(O^l|\mathbf{g}_c^l)P(\mathbf{g}_c^l) = \left(\prod_q^{Q-1} p(O_q^l|g_c^l(q)) \right) \left(\prod_q^{Q-1} P(g_c^l(q)) \right) \quad (3.54)$$

Chaque élément des produits est calculé avec les équations 3.55 et 3.56. La variable $\omega_s^l(f)$ représente la pondération de chaque particule pour la source s à la trame l . La variable $p(O_q^l|(\mathbf{z}_{\mathbf{part}})_{g_r^l(q)}^l(f))$ est définie précédemment dans l'équation 3.41. La variable P_q représente la probabilité que la source potentielle q existe réellement (donc qu'elle ne soit pas une fausse détection). La constante P_{new} représente la probabilité *a priori* qu'une nouvelle source apparaisse (fixée à $P_{new} = 0.005$). Pour sa part, la constante P_{false} représente la probabilité *a priori* qu'une source potentielle soit une fausse détection (fixée à $P_{false} = 0.05$). Finalement, la variable $P(Obs_s^l|\mathbf{O}^{l-1})$ représente la probabilité que la source s soit observée.

$$p(O_q^l|g_c^l(q)) = \begin{cases} 1/4\pi & g_c^l(q) = -2 \\ 1/4\pi & g_c^l(q) = -1 \\ \sum_{f=0}^{F-1} \omega_{g_c^l(q)}^{l-1}(f) p(O_q^l|(\mathbf{z}_{\mathbf{part}})_{g_c^l(q)}^l(f)) & 0 \leq g_c^l(q) < (S-1) \end{cases} \quad (3.55)$$

$$p(g_c^l(q)) = \begin{cases} (1 - P_q)P_{false} & g_c^l(q) = -2 \\ P_q P_{new} & g_c^l(q) = -1 \\ P_q P(Obs_s^l|\mathbf{O}^{l-1}) & 0 \leq g_c^l(q) < (S-1) \end{cases} \quad (3.56)$$

La probabilité qu'une source potentielle q soit une source existante (P_q) est définie à partir de l'énergie du formateur de faisceaux pour la première source potentielle et ensuite est fixe pour les sources potentielles suivantes, telle que définie dans les équations 3.57 et 3.58. La variable E_T représente un seuil d'énergie et il a été déterminé expérimentalement que $E_T = 200$.

$$v_q = \frac{(E_{pot})_q^l}{E_T} \quad (3.57)$$

$$P_q = \begin{cases} v_q^2/2, & q = 0, v_q \leq 1 \\ 1 - v_q^{-2}/2 & q = 0, v_q > 1 \\ 0.3 & q = 1 \\ 0.16 & q = 2 \\ 0.03 & q = 3 \end{cases} \quad (3.58)$$

La probabilité qu'une source soit observée ($P(Obs_s^l | \mathbf{O}^{l-1})$) dépend des probabilités qu'elle existe ($P(E_s^l | \mathbf{O}^{l-1})$) et qu'elle soit active ($P(A_s^l | \mathbf{O}^{l-1})$). Ces probabilités peuvent être calculées récursivement avec les équations 3.59 à 3.63. La variable P_s^{l-1} représente la probabilité que la source s soit déjà suivie. La constante P_o représente la probabilité *a priori* qu'une source ne soit pas observée malgré le fait qu'elle existe (fixée à $P_o = 0.5$). La constante $P(A_s^l | A_s^{l-1})$ représente la probabilité qu'une source active demeure active, tandis que la constante $P(A_s^l | \neg A_s^{l-1})$ représente la probabilité qu'une source active devienne inactive. Pour cette application $P(A_s^l | A_s^{l-1}) = 0.7$ et $P(A_s^l | \neg A_s^{l-1}) = 0.3$. Les constantes P_b et P_m ont des valeurs de 0.15 et 0.85 respectivement. La variable $P(A_s^{l-1} | \mathbf{O}^{l-1})$ représente la probabilité qu'une source suivie s soit active à la trame $l - 1$ selon les observations de la trame $l - 1$ seulement.

$$P(Obs_s^l | \mathbf{O}^{l-1}) = P(E_s^l | \mathbf{O}^{l-1})P(A_s^l | \mathbf{O}^{l-1}) \quad (3.59)$$

$$P(E_s^l | \mathbf{O}^{l-1}) = P_s^{l-1} + (1 - P_s^{l-1}) \left(\frac{P_o P(E_s^{l-1} | \mathbf{O}^{l-2})}{1 - (1 - P_o) P(E_s^{l-1} | \mathbf{O}^{l-2})} \right) \quad (3.60)$$

$$P(A_s^l | \mathbf{O}^{l-1}) = P(A_s^l | A_s^{l-1})P(A_s^{l-1} | \mathbf{O}^{l-1}) + P(A_s^l | \neg A_s^{l-1})(1 - P(A_s^{l-1} | \mathbf{O}^{l-1})) \quad (3.61)$$

$$P(A_s^{l-1} | \mathbf{O}^{l-1}) = \frac{1}{1 + \frac{(1 - P(A_s^{l-1} | \mathbf{O}^{l-2}))(1 - P(A_s^{l-1} | \mathbf{O}^{l-1}))}{P(A_s^{l-1} | \mathbf{O}^{l-2})P(A_s^{l-1} | \mathbf{O}^{l-1})}} \quad (3.62)$$

$$P(A_s^{l-1} | \mathbf{O}^{l-1}) = P_b + P_m P_s^{l-1} \quad (3.63)$$

La probabilité qu'une source s déjà suivie soit associée à une source potentielle (P_s^{l-1}) est calculée d'après l'équation 3.64. La variable $P_s^l(q)$ est calculée à partir de l'équation 3.47. Il s'agit ici de la probabilité pour la trame précédente (d'où l'indice $l - 1$).

$$P_s^{l-1} = \sum_{q=0}^{Q-1} P_s^{l-1}(q) \quad (3.64)$$

3.2.4 Mise à jour de la pondération des particules

La mise à jour de la pondération des particules $(\omega_s^l(f))$ est effectuée à l'aide des équations 3.65 et 3.66. Les variables s , l et f représentent les indices du filtre, de la trame et de la particule respectivement.

$$(\omega_s^l)(f) = \frac{p((\mathbf{z}_{\text{part}})_s^l(f)|O^l)(\omega_s^{l-1})(f)}{\sum_{f=0}^{F-1} p((\mathbf{z}_{\text{part}})_s^l(f)|O^l)(\omega_s^{l-1})(f)} \quad (3.65)$$

$$p((\mathbf{z}_{\text{part}})_s^l(f)|O^l) = (1 - P_s^l) \frac{1}{F} + P_s^l \frac{\sum_{q=0}^{Q-1} P_s^l(q) p(O_q^l | (\mathbf{z}_{\text{part}})_s^l(f))}{\sum_{f=0}^{F-1} \sum_{q=0}^{Q-1} P_s^l(q) p(O_q^l | (\mathbf{z}_{\text{part}})_s^l(f))} \quad (3.66)$$

La probabilité $p(O_q^l | (\mathbf{z}_{\text{part}})_s^l(f))$ est obtenue à partir de l'équation 3.41. Pour sa part, la probabilité P_s^l est obtenue à l'aide de l'équation 3.67. Celle-ci est semblable à l'équation 3.64, à l'exception que cette fois-ci il s'agit de la trame courante d'indice l .

$$P_s^l = \sum_{q=0}^{Q-1} P_s^l(q) \quad (3.67)$$

3.2.5 Ajout d'une nouvelle source

Lorsque la probabilité $P_{H_2}^l(q)$ excède un seuil défini ($T_{\text{new}} = 0.3$), un nouveau filtre est créé à l'indice s_{new} . À ce moment, F particules sont initialisées telles que le décrivent les équations 3.68 et 3.69.

$$(\mathbf{z}_{\text{part}})_{s_{\text{new}}}^l(f) = (\mathbf{z}_{\text{pot}})_q^l \quad 0 \leq f < F \quad (3.68)$$

$$\omega_{s_{\text{new}}}^l(f) = \frac{1}{F} \quad 0 \leq f < F \quad (3.69)$$

L'existence de la source est confirmée lorsque la probabilité d'existence $P(E_s^l | \mathbf{O}^{l-1})$ dépasse un niveau fixe ($T_{\text{conf}} = 0.5$) pendant un nombre minimal de trames ($L_{\text{conf}} = 50$).

3.2.6 Suppression d'une source existante

Lorsque la probabilité qu'une source déjà suivie soit associée à une source potentielle P_s^l est en-dessous d'un seuil défini ($T_{delete} = 0.5$) durant une nombre de trames déterminé ($L_{delete} = 25$), le filtre particulière associé à cette source est supprimé.

3.2.7 Estimation de la position des sources

La position de chaque source suivie est estimée à partir des particules du filtre $(\mathbf{z}_{\text{part}})_s^l(f)$ et de leur pondération respective $\omega_s^l(f)$, telle que donnée par l'équation 3.70. Les variables s , l et f représentent les indices du filtre, de la trame et de la particule respectivement.

$$(\mathbf{z}_{\text{track}})_s^l = \sum_{f=0}^{F-1} \omega_s^l(f) (\mathbf{z}_{\text{part}})_s^l(f) \quad (3.70)$$

3.2.8 Échantillonnage des particules

Si rien n'est fait, les particules d'un filtre finissent par posséder une pondération qui tend vers zéro, tandis qu'une seule particule possède une pondération d'une valeur unitaire. Pour éviter ce phénomène, de nouvelles particules sont échantillonnées à partir des anciennes lorsque la dispersion (E_{disp}) de celles-ci, telle que calculée par l'équation 3.71, est inférieure à un certain seuil ($T_{disp} = 0.7$ dans le cas présent).

$$E_{disp} = \left[\frac{1}{F} \left(\sum_{f=0}^{F-1} [\omega_s^l(f)]^2 \right) \right]^{-1} \quad (3.71)$$

Lorsque cette condition est satisfaite, une variable aléatoire discrète (Ψ_s^l) dont la densité de probabilités est reliée à la pondération des particules est créée en utilisant l'équation 3.72.

$$P(\Psi_s^l = f) = \omega_s^l(f) \quad (3.72)$$

La position des nouvelles particules correspond à la position de l'ancienne particule dont l'indice est relié à une réalisation de la variable aléatoire, telles que l'exposent les équations 3.73 et 3.74.

$$f_{sample} = \Psi_s^l \quad (3.73)$$

$$(\mathbf{z}_{\text{part}}^{\text{new}})_s^l(f) = (\mathbf{z}_{\text{part}})_s^l(f_{sample}) \quad (3.74)$$

3.3 Séparation de sources sonores avec ManyEars

La séparation des sources sonores s'effectue en utilisant leurs positions préalablement obtenues aux cours des étapes de localisation et suivi, tel que l'illustre la figure 3.5.

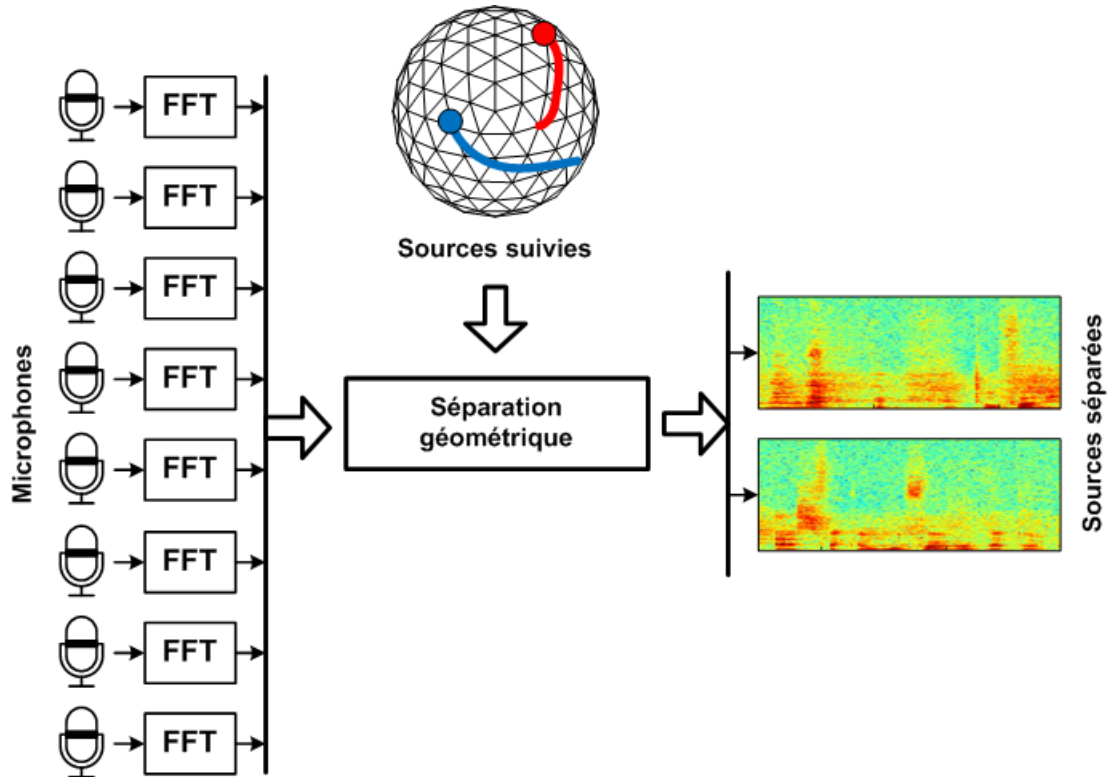


Figure 3.5 Aperçu de la séparation de sources sonores avec ManyEars

Une fois les positions des sources sonores connues, il est possible de séparer ces sources pour isoler le signal de chaque locuteur et ainsi générer des trames indépendantes pour chacun d'entre eux. Pour ce faire, les vecteurs $(\mathbf{X}_{\text{mic}})^l[k]$ et $(\mathbf{X}_{\text{sep}})^l[k]$ sont d'abord définis par les équations 3.75 et 3.76. Le vecteur $(\mathbf{X}_{\text{mic}})^l[k]$ représente les signaux des microphones dans le domaine fréquentiel. Le vecteur $(\mathbf{X}_{\text{sep}})^l[k]$ représente les signaux séparés pour les sources sonores suivies. Les variables l et k représentent les indices de la trame et de la

fréquence discrétisée respectivement.

$$(\mathbf{X}_{\text{mic}})^l[k] = \begin{bmatrix} (X_{\text{mic}})_0^l[k] \\ (X_{\text{mic}})_1^l[k] \\ \vdots \\ (X_{\text{mic}})_{M-1}^l[k] \end{bmatrix} \quad (3.75)$$

$$(\mathbf{X}_{\text{sep}})^l[k] = \begin{bmatrix} (X_{\text{sep}})_0^l[k] \\ (X_{\text{sep}})_1^l[k] \\ \vdots \\ (X_{\text{sep}})_{S-1}^l[k] \end{bmatrix} \quad (3.76)$$

Pour obtenir le signal de chaque source à partir des signaux des microphones, une matrice $\mathbf{W}^l[k]$ de dimension $S \times M$ est utilisée. Cette matrice vise à reproduire l'effet d'un formateur de faisceaux en plus de maximiser la décorrélation entre les signaux des sources suivies, puisque ces derniers sont indépendants. Les équations 3.77 et 3.78 décrivent cette opération.

$$\mathbf{W}^l[k] = \begin{bmatrix} w_{0,0}^l & w_{0,1}^l & \cdots & w_{0,M-1}^l \\ w_{1,0}^l & w_{1,1}^l & \cdots & w_{1,M-1}^l \\ \vdots & \vdots & \ddots & \vdots \\ w_{S-1,0}^l & w_{S-1,1}^l & \cdots & w_{S-1,M-1}^l \end{bmatrix} \quad (3.77)$$

$$(\mathbf{X}_{\text{sep}})^l[k] = \mathbf{W}^l[k](\mathbf{X}_{\text{mic}})^l[k] \quad (3.78)$$

La mise à jour de la matrice $\mathbf{W}^l[k]$ est réalisée par l'équation 3.79.

$$\mathbf{W}^{(l+1)}[k] = (1 - \lambda\mu)\mathbf{W}^l[k] - \mu \left[\alpha^l[k] \frac{\partial J_1(\mathbf{W}^l[k])}{\partial (\mathbf{W}^l)^*[k]} + \frac{\partial J_2(\mathbf{W}^l[k])}{\partial (\mathbf{W}^l)^*[k]} \right] \quad (3.79)$$

La constante μ représente le taux d'adaptation. Ce taux doit être sélectionné de façon conservatrice car une valeur trop élevée peut rendre le système instable. Pour l'application courante, $\mu = 0.002$. La constante λ représente le taux de régularisation. La première contrainte $J_1(\mathbf{W}^l[k])$, obtenue par l'équation 3.80, vise à maximiser la décorrélation des sources. Le gradient de cette contrainte $\frac{\partial J_1(\mathbf{W}^l[k])}{\partial (\mathbf{W}^l)^*[k]}$ est évalué grâce à l'équation 3.82. L'expression $(.)^H$ représente la matrice adjointe. La seconde contrainte $J_2(\mathbf{W}^l[k])$ (équation 3.81) vise à maximiser la décorrélation entre les signaux des sources suivies.

tion 3.81) est utilisée pour garantir un gain unitaire dans la direction de la source suivie et un gain nul dans les autres directions. Le gradient de cette contrainte $\frac{\partial J_2(\mathbf{W}^l[k])}{\partial (\mathbf{W}^l)^*[k]}$ est évalué dans l'équation 3.83. Finalement, l'énergie est normalisée grâce au facteur $\alpha^l[k]$ évalué dans l'équation 3.84.

$$J_1(\mathbf{W}^l[k]) = \|\mathbf{R}_{ss}[k] - \text{diag}[\mathbf{R}_{ss}[k]]\|^2 \quad (3.80)$$

$$J_2(\mathbf{W}^l[k]) = \|\mathbf{W}^l[k]\mathbf{A}^l[k] - \mathbf{I}\|^2 \quad (3.81)$$

$$\frac{\partial J_1(\mathbf{W}^l[k])}{\partial (\mathbf{W}^l)^*[k]} = 4 \left(\mathbf{E}^l[k]\mathbf{W}^l[k](\mathbf{X}_{\text{mic}})^l[k] \right) (\mathbf{X}_{\text{mic}})^l[k]^H \quad (3.82)$$

$$\frac{\partial J_2(\mathbf{W}^l[k])}{\partial (\mathbf{W}^l)^*[k]} = 2[\mathbf{W}^l[k]\mathbf{A}^l[k] - \mathbf{I}]\mathbf{A}^l[k]^H \quad (3.83)$$

$$\alpha^l[k] = \|\mathbf{R}_{mm}^l[k]\|^{-2} \quad (3.84)$$

La matrice $\mathbf{A}^l[k]$ représente le délai d'arrivée du signal de chaque source sonore selon la position de la source. Elle est définie dans les équations 3.85, 3.86 et 3.87.

$$\mathbf{A}^l[k] = \begin{bmatrix} a_{0,0}^l & a_{0,1}^l & \cdots & a_{0,S-1}^l \\ a_{1,0}^l & a_{1,1}^l & \cdots & a_{1,S-1}^l \\ \vdots & \vdots & \ddots & \vdots \\ a_{M-1,0}^l & a_{M-1,1}^l & \cdots & a_{M-1,S-1}^l \end{bmatrix} \quad (3.85)$$

$$a_{m,s}^l[k] = e^{-j2\pi k\tau_{m,s}^l} \quad (3.86)$$

$$\tau_{m,s}^l = \text{round} \left[\left(\frac{F_s}{c_{air}} \right) \|(\mathbf{z}_{\text{track}})^l_s - (\mathbf{z}_{\text{mic}})_m\| \right] \quad (3.87)$$

Les matrices de corrélation pour les signaux des microphones et des sources séparées sont estimées telles que données par les équations 3.88 et 3.89.

$$\mathbf{R}_{\text{mm}}^l[k] = (\mathbf{X}_{\text{mic}})^l[k](\mathbf{X}_{\text{mic}})^l[k]^H \quad (3.88)$$

$$\mathbf{R}_{\text{ss}}^l[k] = (\mathbf{X}_{\text{sep}})^l[k](\mathbf{X}_{\text{sep}})^l[k]^H \quad (3.89)$$

L'expression intermédiaire $\mathbf{E}^l[k]$ est également calculée dans l'équation 3.90.

$$\mathbf{E}^l[k] = \mathbf{R}_{\text{ss}}^l[k] - \text{diag}(\mathbf{R}_{\text{ss}}^l[k]) \quad (3.90)$$

Lorsqu'une nouvelle source est suivie, la ligne de la matrice $\mathbf{W}^l[k]$ associée à cette nouvelle source est initialisée avec les valeurs de la matrice $\mathbf{A}^l[k]$ à l'aide de l'équation 3.91.

$$\mathbf{W}^l[k] = \begin{bmatrix} w_{0,0}^l & w_{0,1}^l & \cdots & w_{0,M-1}^l \\ w_{1,0}^l & w_{1,1}^l & \cdots & w_{1,M-1}^l \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(a_{S-1,0}^l)^*}{M} & \frac{(a_{S-1,1}^l)^*}{M} & \cdots & \frac{(a_{S-1,M-1}^l)^*}{M} \end{bmatrix} \quad (3.91)$$

3.4 Post-filtrage sur ManyEars

Le post-filtrage est utilisé pour rehausser le rapport signal sur bruit d'une source déterminée afin de réduire la corruption du signal par le bruit stationnaire, la réverbération et les autres sources sonores présentes pour faciliter la reconnaissance. Chaque source séparée est post-filtrée, tel qu'illustré dans la figure 3.6. Le signal post-filtré est moins bruité mais présente certains effets musicaux causés par le changement de pondération des bandes de fréquences.

Pour ce faire, un gain est appliqué sur chaque bande de fréquences tel que donné par l'équation 3.92. Les variables l et k représentent les indices de la trame et de la fréquence discrétisée respectivement.

$$(X_{\text{post}})_s^l[k] = G_s^l[k](X_{\text{sep}})_s^l[k] \quad (3.92)$$

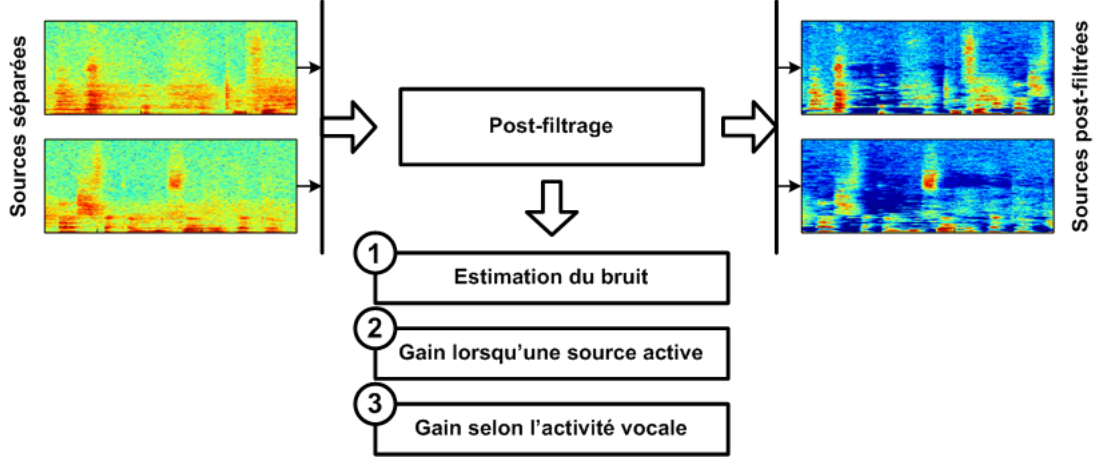


Figure 3.6 Aperçu du post-filtrage de sources sonores avec ManyEars

3.4.1 Estimation du bruit

Le bruit stationnaire de l'environnement $(\lambda_{sep}^{stat})_s^l[k]$ est d'abord estimé avec la même méthode MCRA donnée à la section 3.1. Les équations 3.12 à 3.18 sont utilisées et demeurent identiques, à l'exception du signal de la source s séparée $(X_{sep})_s^l[k]$ qui remplace le signal $(X_{mic})_m^l[k]$ du microphone m . Les variables l et k représentent les indices de la trame et de la fréquence discrétisée respectivement.

Pour cette application, la réverbération n'est pas prise en compte car il a été observé expérimentalement qu'elle diminuait les performances du système dans le cas où une reconnaissance vocale était effectuée par la suite. Quant au bruit provenant des autres sources sonores actives, il est estimé à l'aide des équations 3.93 et 3.94.

$$Z_s^l[k] = \alpha_t Z_s^{l-1}[k] + (1 - \alpha_t) |(X_{sep})_s^l[k]|^2 \quad (3.93)$$

$$(\lambda_{sep}^{leak})_s^l[k] = \eta \sum_{s'=0, s' \neq s}^{S-1} Z_{s'}^l[k] \quad (3.94)$$

La constante α_t est utilisée pour lisser le spectre ($\alpha_t = 0.1$ pour cette application). Puisque les autres sources sonores sont atténuées durant l'étape de séparation, un facteur η est utilisé pour diminuer l'amplitude de ces sources ($\eta = 0.3$). Le bruit total est finalement calculé à l'aide de l'équation 3.95.

$$(\lambda_{sep})_s^l[k] = (\lambda_{sep}^{stat})_s^l[k] + (\lambda_{sep}^{leak})_s^l[k] \quad (3.95)$$

3.4.2 Gain lorsqu'une source est active

Le gain $(G_{H_1})_s^l[k]$ à appliquer lorsqu'une source sonore est active (le locuteur n'est pas dans une période de silence) est dérivé dans cette section. Le rapport signal sur bruit *a posteriori* $\gamma_s^l[k]$ est défini par l'équation 3.96. Les variables s , l et k représentent les indices de la source séparée, de la trame et de la fréquence discrétisée respectivement.

$$(\gamma_{sep})_s^l[k] = \frac{|(X_{sep})_s^l[k]|^2}{(\lambda_{sep})_s^l[k]} \quad (3.96)$$

Le rapport signal sur bruit *a priori* $(\xi_{sep})_s^l[k]$ est estimé par l'équation 3.97.

$$(\xi_{sep})_s^l[k] = (1 - (\alpha_{sep}^p)_s^l[k])(G_{H_1}^2)_s^{l-1}[k](\gamma_{sep})_s^l[k] + (\alpha_{sep}^p)_s^l[k] \max\{[(\gamma_{sep})_s^l[k] - 1], 0\} \quad (3.97)$$

Le taux d'adaptation $(\alpha_{sep}^p)_s^l[k]$ est évalué par l'équation 3.98. Pour ce système, $\alpha_{pmin} = 0.07$.

$$(\alpha_{sep}^p)_s^l[k] = \left(\frac{(\xi_{sep})_s^l[k]}{(\xi_{sep})_s^l[k] + 1} \right)^2 + \alpha_{pmin} \quad (3.98)$$

Le gain $(G_{H_1})_s^l[k]$ est ensuite calculé selon l'équation 3.99.

$$(G_{H_1})_s^l[k] = \left(\frac{(\xi_{sep})_s^l[k]}{(\xi_{sep})_s^l[k] + 1} \right) ((G_t)_s^l[k]) \quad (3.99)$$

Le gain $(G_t)_s^l[k]$ est normalement représenté par une fonction transcendante. Dans le cas présent, on l'estime par une fonction simplifiée qui est décrite aux équation 3.100 à 3.105.

$$v_s^l[k] = \left(\frac{(\xi_{sep})_s^l[k]}{(\xi_{sep})_s^l[k] + 1} \right) [(\gamma_{sep})_s^l[k]] \quad (3.100)$$

$$(G_t)_s^l[k] = \begin{cases} (1 - (\chi_{frac})_s^l[k])(\phi_{int})_s^l[k] + (\chi_{frac})_s^l[k](\phi_{frac})_s^l[k] & v_s^l[k] \leq 9.5 \\ 1 & v_s^l[k] > 9.5 \end{cases} \quad (3.101)$$

$$(\phi_{int})_s^l[k] = \kappa[(\chi_{int})_s^l[k]] \quad (3.102)$$

$$(\phi_{frac})_s^l[k] = \frac{\kappa[(\chi_{int})_s^l[k] + 1]}{\sqrt{v_s^l[k] + \frac{1}{10000}}} \quad (3.103)$$

$$(\chi_{int})_s^l[k] = \text{floor}(2v_s^l[k]) \quad (3.104)$$

$$(\chi_{frac})_s^l[k] = 2v_s^l[k] - (\chi_{int})_s^l[k] \quad (3.105)$$

La variable κ est un vecteur de 21 éléments dont l'indice commence à 0 et se termine à 20 tel qu'exposé par l'équation 3.106. Ce vecteur est mis en mémoire et permet d'évaluer rapidement l'équation 3.101. Le logarithme naturel du gain est illustré à la figure 3.7.

$$\kappa = \begin{bmatrix} 0.75008, & 0.93640, & 1.11688, & 1.28852, & 1.45013, & 1.60184, & 1.74422, & \dots \\ 1.87812, & 2.00448, & 2.12417, & 2.23799, & 2.34656, & 2.45053, & 2.55031, & \dots \\ 2.64646, & 2.73921, & 2.82895, & 2.91602, & 3.00049, & 3.08265, & 3.16270 \end{bmatrix} \quad (3.106)$$

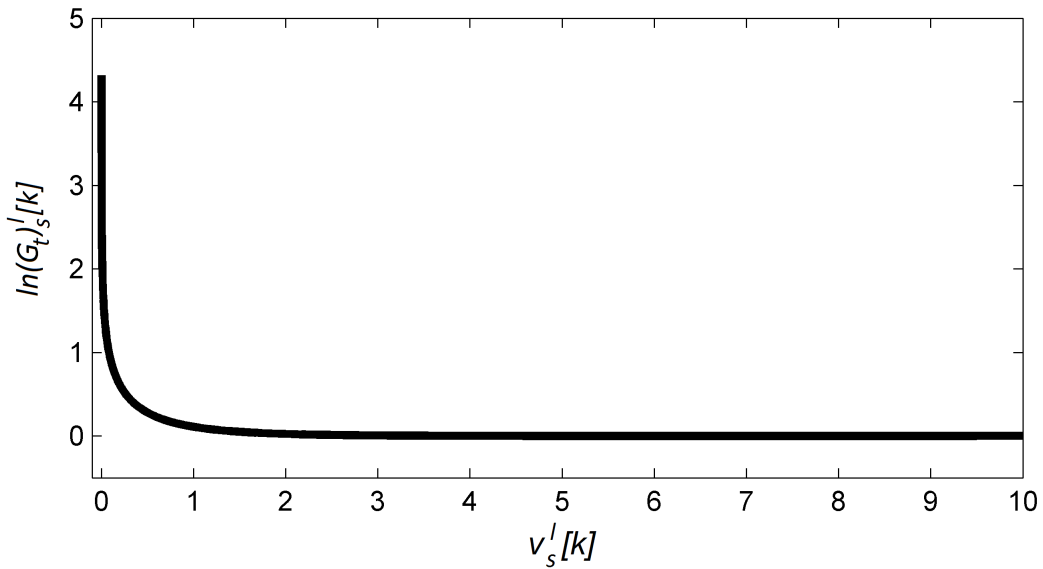


Figure 3.7 Fonction transcendante du gain pour une source active

3.4.3 Gain selon l'activité vocale

Dans la section 3.4.2, le gain est optimisé lorsqu'une source sonore est active. Un gain minimal ($G_{min} = 0.1$) doit également être présent lorsque cette source n'est pas active. Le gain final décrit par l'équation 3.107 est donc un combinaison de ces deux gains et sa valeur dépend de la probabilité que la source sonore soit active $((p_{speech})_s^l[k])$. Les variables s , l et k représentent les indices de la source séparée, de la trame et de la fréquence discrétisée respectivement.

$$G_s^l[k] = ((G_{H_1})_s^l[k])^{(p_{speech})_s^l[k]} (G_{min})^{(1-(p_{speech})_s^l[k])} \quad (3.107)$$

La probabilité qu'une source sonore soit active est évaluée avec les équations 3.108 et 3.109.

$$(p_{speech})_s^l[k] = \left[1 + \left(\frac{(q_{speech})_s^l[k]}{1 - (q_{speech})_s^l[k]} \right) (1 + (\xi_{sep})_s^l[k]) e^{-v_s^l[k]} \right]^{-1} \quad (3.108)$$

$$(q_{speech})_s^l[k] = \min (1 - (P_{local})_s^l[k] (P_{global})_s^l[k] (P_{frame})_s^l[k], 0.9) \quad (3.109)$$

Les probabilités d'activité vocale sur une fenêtre étroite, une fenêtre large et une fenêtre entière sont calculées avec les équations 3.110 et 3.111. La variable ψ est remplacée par *local*, *global* et *frame* pour chaque type de fenêtre. Pour cette application, $\theta = 0.5$ et $\alpha_\zeta = 0.3$. Lorsque l'indice de $(\xi_{sep})_s^l$ est plus petit que 0 ou égal ou supérieur à N , une valeur de 0 est retournée.

$$(P_\psi)_s^l[k] = \frac{1}{1 + \left(\frac{\theta}{(\zeta_\psi)_s^l[k]} \right)^2} \quad (3.110)$$

$$(\zeta_\psi)_s^l[k] = (1 - \alpha_\zeta) (\zeta_\psi)_s^{l-1}[k] + \alpha_\zeta \sum_{c=-\omega_\psi}^{\omega_\psi} h_\psi[c] (\xi_{sep})_s^l[k + c] \quad (3.111)$$

La fenêtre de filtrage Hann utilisée dans l'équation 3.111 est décrite par l'équation 3.112. Quant aux dimensions des fenêtres étroite, large et entière, elles sont données au tableau 3.5.

$$h_\psi[c_\psi] = 0.5 \left(1 - \cos \left(\frac{\pi(c_\psi + \omega_\psi + 1)}{\omega_\psi + 1} \right) \right) \quad -\omega_\psi \leq c_\psi \leq \omega_\psi \quad (3.112)$$

Tableau 3.5 Fenêtres utilisées pour la détection de l'activité vocale

Type	ψ	ω_ψ
Fenêtre étroite	<i>local</i>	1
Fenêtre large	<i>global</i>	15
Fenêtre entière	<i>frame</i>	511

Ce post-filtrage permet d'effectuer une reconnaissance de la parole durant un dialogue entre plusieurs locuteurs [46].

CHAPITRE 4

WISS, UN SYSTÈME DE RECONNAISSANCE DE LOCUTEURS

Bien que ManyEars permet de séparer simultanément plusieurs sources sonores, le système WISS conçu effectue une reconnaissance de locuteurs pour un locuteur unique (et non pour plusieurs locuteurs simultanément). Ce système devait initialement permettre au robot de reconnaître plusieurs locuteurs qui parlent simultanément. Cependant, le faible SNR observé dans l'environnement au sein duquel le robot interagit introduit la nécessité d'estimer les bruits additif et convolutif. La présence de bruit non-stationnaire en provenance des autres locuteurs vient perturber cette estimation, ce qui empêche une interaction à plusieurs locuteurs. La séparation de plusieurs locuteurs du système ManyEars est donc superflue pour l'application courante. Par contre, le choix du système ManyEars reste pertinent car il permet d'effectuer le prétraitement pour une reconnaissance de la parole avec plusieurs personnes, et une reconnaissance de locuteur pour une personne avec WISS. La reconnaissance s'effectue en deux étapes : l'entraînement des modèles et la comparaison des caractéristiques vocales avec les modèles préalablement entraînés. Contrairement à la deuxième étape, l'entraînement doit s'effectuer dans un environnement peu bruyant afin de modéliser fidèlement chaque locuteur.

Le système WISS comporte deux niveaux de traitement. Le premier niveau consiste à extraire les caractéristiques vocales à partir du signal en provenance du locuteur. Le second niveau consiste à comparer ces caractéristiques à des modèles entraînés précédemment pour déterminer l'identité du locuteur actuel.

L'environnement dans lequel le robot tente d'identifier un locuteur comporte de nombreuses sources de perturbations. Voici les hypothèses établies pour un scénario typique, définissant le modèle utilisé pour les bruits convolutif et additif, le tout schématisé à la figure 4.1 :

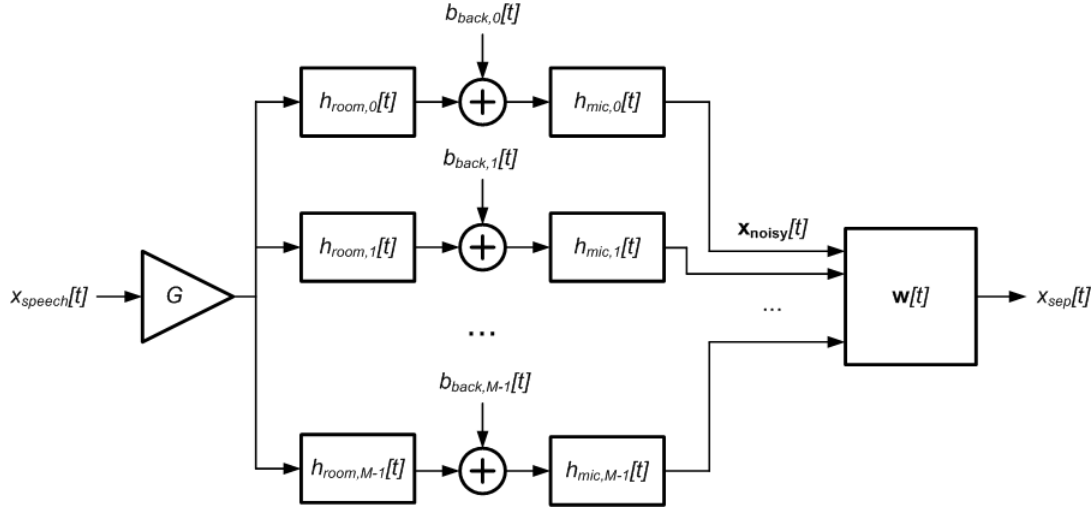


Figure 4.1 Modélisation des bruits convolutif et additif

- **Signal du locuteur** : Le signal $x_{speech}[t]$ est considéré comme un processus aléatoire.
- **Volume du locuteur** : Le locuteur parle à un volume inconnu. Ce volume est représenté par un gain G .
- **Réverbération de la pièce** : Une réverbération inconnue est présente entre le locuteur et chaque microphone m . Elle dépend des positions du locuteur et du robot, et des caractéristiques physiques de la pièce. Le locuteur et le robot se déplacent lentement durant le dialogue, de sorte que la réverbération demeure peu changée sur une période de quelques secondes. La réverbération est modélisée par un filtre de réponse impulsionnelle $h_{room,m}[t]$.
- **Bruit ambiant** : Un bruit de fond est présent dans la pièce et il est habituellement causé par le système de ventilation et par le fonctionnement de certains appareils électriques. Ce bruit est coloré et généralement rose, ce qui signifie que l'énergie de la densité de puissance se situe en majorité dans les basses fréquences. Chaque microphone m capte son propre bruit de fond $b_{back,m}[t]$.
- **Réponse en fréquence des microphones** : La réponse en fréquence des microphones est inconnue. On suppose qu'elle demeure inchangée dans le temps et que les microphones sont linéaires en amplitude, donc qu'ils ne sont pas sujets à des distortions causées par un changement d'amplitude du signal à l'entrée. Cette hypothèse est raisonnable car le volume du locuteur demeure dans un interval limité. Chaque microphone m est représenté par un filtre ayant pour réponse impulsionnelle $h_{mic,m}[t]$.
- **Indépendance du signal et du bruit** : Les processus aléatoires $x_{speech}[t]$ et $b_{back,m}[t]$ sont indépendants.

Le signal perçu à l'entrée de chaque microphone m est représenté par la variable $(x_{noisy})_m[t]$. Les signaux de tous les microphones sont regroupés au sein du vecteur $\mathbf{x}_{noisy}[t]$, tels que montrés dans l'équation 4.1.

$$\mathbf{x}_{noisy}[t] = \begin{bmatrix} (x_{noisy})_0[t] \\ (x_{noisy})_1[t] \\ \dots \\ (x_{noisy})_{M-1}[t] \end{bmatrix} \quad (4.1)$$

Selon les hypothèses précédentes, les signaux perçus par les microphones peuvent être reliés au signal du locuteur par l'équation 4.2. Le symbole $(*)$ représente la convolution linéaire.

$$\mathbf{x}_{noisy}[t] = \begin{bmatrix} \{(Gx_{speech}[t]) * (h_{room,0}[t]) + b_{back,0}[t]\} * h_{mic,0}[t] \\ \{(Gx_{speech}[t]) * (h_{room,1}[t]) + b_{back,1}[t]\} * h_{mic,1}[t] \\ \dots \\ \{(Gx_{speech}[t]) * (h_{room,M-1}[t]) + b_{back,M-1}[t]\} * h_{mic,M-1}[t] \end{bmatrix} \quad (4.2)$$

Cette notation peut être simplifiée sous sa forme vectorielle par les équations 4.3, 4.4 et 4.5.

$$\mathbf{x}_{noisy}[t] = \mathbf{h}_{noisy}[t] * x_{speech}[t] + \mathbf{b}_{noisy}[t] \quad (4.3)$$

$$\mathbf{h}_{noisy}[t] = \begin{bmatrix} Gh_{room,0}[t] * h_{mic,0}[t] \\ Gh_{room,1}[t] * h_{mic,1}[t] \\ \dots \\ Gh_{room,M-1}[t] * h_{mic,M-1}[t] \end{bmatrix} \quad (4.4)$$

$$\mathbf{b}_{noisy}[t] = \begin{bmatrix} b_{back,0}[t] * h_{mic,0}[t] \\ b_{back,1}[t] * h_{mic,1}[t] \\ \dots \\ b_{back,M-1}[t] * h_{mic,M-1}[t] \end{bmatrix} \quad (4.5)$$

L'équation 3.78 indique que le signal séparé est obtenu en multipliant le vecteur des microphones $(\mathbf{X}_{mic})^l[k]$ par une matrice de décorrélation $\mathbf{W}^l[k]$. Ceci est l'équivalent d'une convolution dans le domaine temporel, tel qu'illustré par l'équation 4.6.

$$x_{sep}[t] = \mathbf{w}[t] * \mathbf{x}_{noisy}[t] \quad (4.6)$$

Comme l'indiquent les équations 4.7 à 4.10, cette matrice de décorrélation $\mathbf{w}[t]$, une fois multipliée par le vecteur $\mathbf{x}_{noisy}[t]$, donne un résultat scalaire.

$$x_{sep}[t] = \mathbf{w}[t] * (\mathbf{h}_{noisy}[t] * x_{speech}[t] + \mathbf{b}_{noisy}[t]) \quad (4.7)$$

$$h[t] = \mathbf{w}[t] * \mathbf{h}_{noisy}[t] \quad (4.8)$$

$$b[t] = \mathbf{w}[t] * \mathbf{b}_{noisy}[t] \quad (4.9)$$

$$x_{sep}[t] = h[t] * x_{speech}[t] + b[t] \quad (4.10)$$

La densité spectrale du processus aléatoire $x_{sep}[n]$ consiste à calculer la transformée de Fourier de son autocorrélation en utilisant l'équation 4.11.

$$|X_{sep}(j\omega)|^2 = \int_{\omega=-\infty}^{\omega=+\infty} E[x_{sep}[t]x_{sep}[t-\tau]]e^{-j\omega\tau}d\tau \quad (4.11)$$

Il est possible d'obtenir par l'équation 4.15 une expression finale pour la densité spectrale. Les différentes étapes pour y parvenir sont décrites par les équations 4.12, 4.13 et 4.14. L'opérateur $E[.]$ représente l'espérance mathématique.

$$|X_{sep}(j\omega)|^2 = \int_{\omega=-\infty}^{\omega=+\infty} \left(\begin{array}{cc} E[& (h[t] * x_{speech}[t] + b[t]) \\ & (h[t-\tau] * x_{speech}[t-\tau] + b[t-\tau]) \end{array} \begin{array}{c} \times \\]e^{-j\omega\tau}d\tau \end{array} \right) \quad (4.12)$$

$$|X_{sep}(j\omega)|^2 = \int_{\omega=-\infty}^{\omega=+\infty} \left(\begin{array}{c} \{ \begin{array}{l} E[(h[t] * x_{speech}[n])(h[t-\tau] * x_{speech}[t-\tau])] + \\ E[(h[t] * x_{speech}[n])(b[t-\tau])] + \\ E[(h[t-\tau] * x_{speech}[t-\tau])(b[t])] + \\ E[(b[t])(b[t-\tau])] \end{array} \} e^{-j\omega\tau}d\tau \end{array} \right) \quad (4.13)$$

$$|X_{sep}(j\omega)|^2 = \int_{\omega=-\infty}^{\omega=+\infty} \left(\frac{\{E[(h[t] * x_{speech}[t])(h[t - \tau] * x_{speech}[t - \tau])] + E[(b[t])(b[t - \tau])]\}}{e^{-j\omega\tau}} d\tau \right) \quad (4.14)$$

$$|X_{sep}(j\omega)|^2 = |H(j\omega)|^2 |X_{speech}(j\omega)|^2 + |B(j\omega)|^2 \quad (4.15)$$

Cette relation démontre donc clairement que, selon les hypothèses émises, la densité spectrale du signal séparé est équivalente à une multiplication du signal d'intérêt par la réponse fréquentielle du canal et l'addition de la densité spectrale du bruit de l'environnement. Le défi revient donc à effectuer une reconnaissance de locuteurs sans que les performances soient affectées par les disparités entre deux environnements modélisés par $|H(j\omega)|^2$ et $|B(j\omega)|^2$.

4.1 Entraînement de WISS

L'entraînement s'effectue à l'aide d'un seul microphone dans un environnement silencieux. Les étapes qui mènent à l'obtention d'un modèle pour le locuteur entraîné sont illustrées à la figure 4.2. Les paramètres présentés dans cette section sont déterminés empiriquement en simulation pour optimiser les performances.

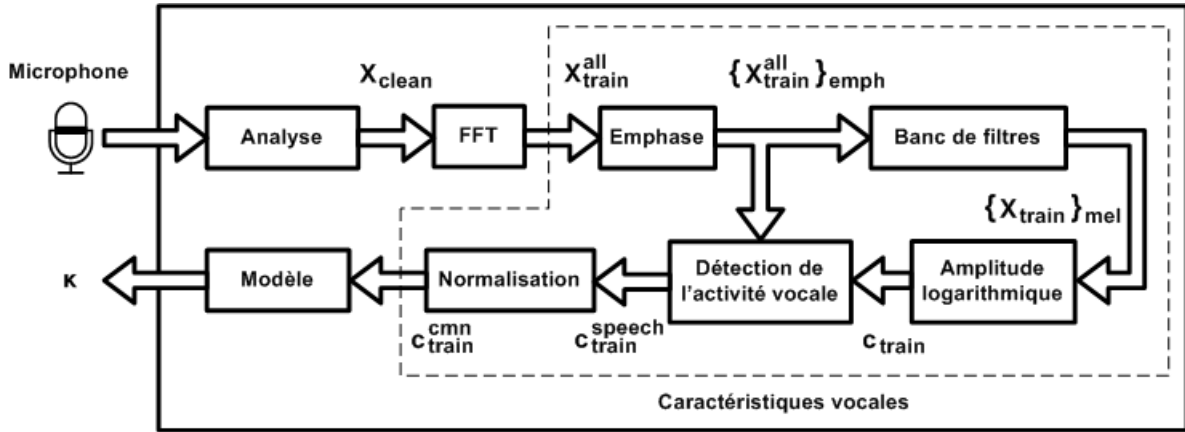


Figure 4.2 Schéma-bloc des modules pour l'entraînement de WISS

4.1.1 Analyse et FFT

La première étape du processus d'entraînement consiste à extraire le spectre des signaux des sources sonores. Pour l'entraînement du locuteur u sur un total de U locuteurs, un signal sans bruit $(x_{clean})_u^l[n]$ est utilisé (enregistré à l'aide d'un microphone sur un poste de

travail dans un environnement silencieux). Les variables l , u et n représentent les indices de la trame, du locuteur et de l'échantillon dans le domaine du temps respectivement. Tel que présenté à l'équation 4.16, le signal est d'abord multiplié par une fenêtre de lissage. Il s'agit de la même fenêtre que celle présentée à l'équation 3.5. Tout comme pour le cas précédent, $N = 1024$ et $o = 0.5$.

$$(x_{clean})_u^l[n] = w[n](x_{clean})_u[n + (l)(N)(o)] \quad 0 \leq n < N, \quad 0 \leq u < U \quad (4.16)$$

Une transformée de Fourier discrète est ensuite appliquée au signal lissé en utilisant l'équation 4.17.

$$(X_{clean})_u^l[k] = \sum_{n=0}^{N-1} (x_{clean})_u^l[n] e^{-j2\pi nk} \quad 0 \leq k < N \quad (4.17)$$

Finalement, la variable $(X_{train}^{all})_u^l[k]$ représente la puissance du spectre pour le nombre total de trames générées (L_{train}^{all}) .

$$(X_{train}^{all})_u^l[k] = |(X_{clean})_u^l[k]|^2 \quad 0 \leq l < (L_{train}^{all})_u \quad (4.18)$$

4.1.2 Caractéristiques vocales

Les caractéristiques utilisées pour la présente application sont similaires à celles du MFCC, à l'exception qu'elles demeurent dans le domaine spectral pour permettre l'utilisation de masques. De plus, pour cette application, seuls les coefficients statiques sont utilisés. Ce choix est justifié par la complexité additionnelle requise pour ces coefficients lors de la modification dynamique des modèles en fonction du bruit stationnaire.

Emphase

Une première étape de filtrage est d'abord appliquée pour mettre l'accent sur les fréquences les plus élevées. Ceci permet d'augmenter l'accent des formants en hautes fréquences qui sont naturellement atténués par rapport à ceux en basses fréquences. Pour la présente application, cette opération est particulièrement utile car elle réduit également la contribution du bruit rose qui se situe dans les basses fréquences. En général, un filtre à réponse impulsionnelle finie est appliqué avant même la fenêtre de lissage. Ce filtre est donné à l'équation 4.19. La constante α_{emph} possède généralement une valeur entre 0.9 et 1.0 [26], qui est fixée à 0.95 dans le cas présent.

$$y[n] = x[n] - \alpha_{emph} x[n-1] \quad (4.19)$$

Puisque le signal obtenu à partir du système ManyEars est déjà disponible dans le domaine spectral, le filtre précédent est appliqué directement dans le domaine fréquentiel. La transformée en Z de la réponse impulsionnelle de ce filtre est donnée à l'équation 4.20.

$$H_{emph}(z) = 1 - \alpha_{emph} z^{-1} \quad (4.20)$$

La réponse en fréquences en terme de puissance de ce filtre est donc obtenue à l'équation 4.21.

$$|H_{emph}(e^{j\omega})|^2 = 1 + (\alpha_{emph})^2 - 2\alpha_{emph} \cos(\omega) \quad (4.21)$$

Dans le cas présent, cette fonction est discrétisée pour chaque indice k , comme le montre l'équation 4.22.

$$H_{emph}[k] = 1 + (\alpha_{emph})^2 - 2\alpha_{emph} \cos\left(\frac{2\pi k}{N}\right) \quad (4.22)$$

Le signal avec emphase sur les hautes fréquences est donc obtenu par l'équation 4.23.

$$(\{X_{train}^{all}\}_{emph})_u^l[k] = H_{emph}[k](X_{train}^{all})_u^l[k] \quad (4.23)$$

Un exemple du logarithme du signal $(\{X_{train}^{all}\}_{emph})_u^l[k]$ est illustré à la figure 4.3. Le logarithme est utilisé afin de faciliter la visualisation. Il est possible de voir à plusieurs temps différents des agglomérations d'énergie qui représentent différentes phonèmes.

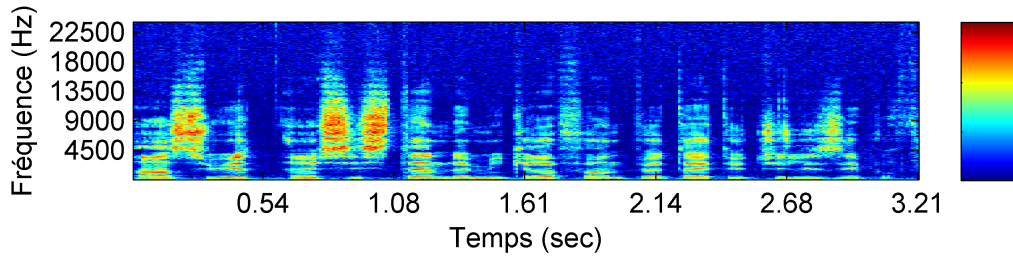


Figure 4.3 Exemple du logarithme du spectre avec emphase pour l'entraînement

Détection de l'activité vocale

Certaines trames représentent des périodes de silence. Puisque celles-ci ne contiennent pas d'information sur le locuteur, seules les trames qui contiennent des segments de parole sont utilisées. Les équations 4.24 et 4.25 servent à déterminer la probabilité que la parole soit présente pour cette section du spectre. La fenêtre $h_\psi[c]$ est la même que celle définie à l'équation 3.112. La variable ψ est également remplacée par *local*, *global* et *frame* pour chaque type de fenêtre. Les dimensions de chaque fenêtre sont présentées au tableau 3.5. Pour notre implémentation, $\alpha_\zeta = 0.3$.

$$(P_\psi)_u^l[k] = \frac{1}{1 + \left(\frac{\theta_u}{(\zeta_\psi)_u^l[k]}\right)^2} \quad (4.24)$$

$$(\zeta_\psi)_u^l[k] = (1 - \alpha_\zeta)(\zeta_\psi)_u^{l-1}[k] + \alpha_\zeta \sum_{c=-\omega_\psi}^{\omega_\psi} h_\psi[c](\{X_{train}^{all}\}_{emph})_u^l[k+c] \quad (4.25)$$

La variable θ_u est obtenue par l'équation 4.26. Il s'agit en fait de déterminer la puissance du signal et de multiplier cette valeur par un facteur θ_{scale} (dans le cas présent, $\theta_{scale} = 3$).

$$\theta_u = \theta_{scale} \left[\frac{1}{(L_{train}^{all})_u} \sum_{l=0}^{(L_{train}^{all})_u-1} \left(\frac{1}{N} \left(\sum_{k=0}^{N-1} (\{X_{train}^{all}\}_{emph})_u^l[k] \right) \right) \right] \quad (4.26)$$

L'activité vocale est ensuite déterminée d'après l'équation 4.27. Le seuil T_{vad} est fixé à une valeur de 10 pour cette application.

$$(P_{vad})_u^l = \begin{cases} 0 & \sum_{k=0}^{N-1} (P_{local})_u^l[k](P_{global})_u^l[k](P_{frame})_u^l[k] < T_{vad} \\ 1 & \sum_{k=0}^{N-1} (P_{local})_u^l[k](P_{global})_u^l[k](P_{frame})_u^l[k] \geq T_{vad} \end{cases} \quad (4.27)$$

Un exemple de l'activité vocale obtenue est illustré à la figure 4.4. Les zones en blanc représentent les trames actives et les zones en noir celles inactives. En comparant avec la figure 4.3, il est observé que les trames possédant peu d'énergie sont considérées comme inactives.

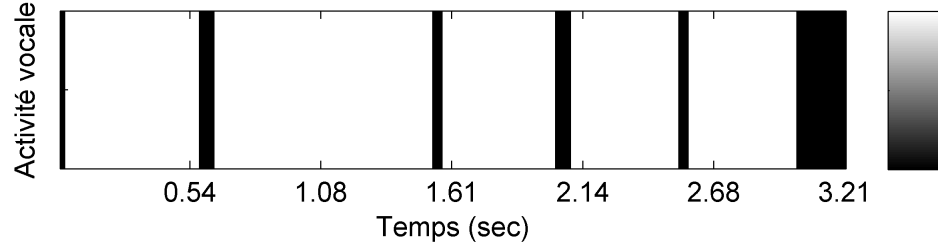


Figure 4.4 Exemple de l'activité vocale pour l'entraînement

Banc de filtres

L'oreille humaine discrimine les fréquences d'une manière logarithmique et par conséquent elle distingue plus facilement les basses fréquences que les hautes fréquences. Un banc de filtres est donc utilisé pour reproduire cet effet. Chaque filtre Λ du banc possède la forme décrite par l'équation 4.28. En général, le nombre de filtres varie entre 24 et 40 (réf. : chapitre 6 [23]). Pour notre implémentation, Λ_{max} est fixé à 24 filtres (cette configuration a démontré de bonnes performances et limite le nombre de dimensions pour chaque vecteur). Les plages de fréquences sont listées au tableau 4.1. Ce banc de filtres est illustré à la figure 4.5. On y observe une largeur de bande qui augmente vers les hautes fréquences.

Tableau 4.1 Banc de filtres

Λ	$(b_{min})_{\Lambda}$	$(b_{center})_{\Lambda}$	$(b_{max})_{\Lambda}$	Λ	$(b_{min})_{\Lambda}$	$(b_{center})_{\Lambda}$	$(b_{max})_{\Lambda}$
0	0	50	100	12	1700	1850	2000
1	100	150	200	13	1980	2150	2320
2	200	250	300	14	2300	2500	2700
3	300	350	400	15	2650	2900	3150
4	390	450	510	16	3100	3400	3700
5	510	570	640	17	3600	4000	4150
6	630	700	770	18	4300	4800	5000
7	760	840	920	19	5200	5800	6200
8	920	1000	1080	20	6300	7000	7700
9	1070	1170	1270	21	7500	8500	9500
10	1260	1370	1480	22	9000	10500	12000
11	1480	1600	1720	23	11500	13500	15500

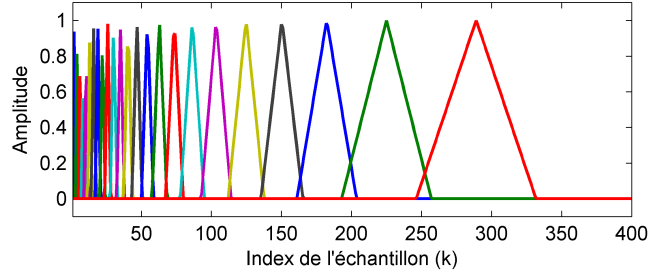


Figure 4.5 Banc de filtres

$$b_{\Lambda}[k] = \begin{cases} 0 & k < \frac{N}{F_s}(b_{min})_{\Lambda} \\ [k(F_s/N) - (b_{min})_{\Lambda}] / [(b_{center})_{\Lambda} - \frac{N}{F_s}(b_{min})_{\Lambda}] & (b_{min})_{\Lambda} \leq k \leq \frac{N}{F_s}(b_{center})_{\Lambda} \\ [k(F_s/N) - (b_{center})_{\Lambda}] / [(b_{center})_{\Lambda} - (b_{max})_{\Lambda}] + 1 & \frac{N}{F_s}(b_{center})_{\Lambda} < k \leq \frac{N}{F_s}(b_{max})_{\Lambda} \\ 0 & k > \frac{N}{F_s}(b_{max})_{\Lambda} \end{cases} \quad (4.28)$$

Le spectre du signal avec emphase est ensuite multiplié par chacun des filtres du banc. La somme des produits est calculée selon l'équation 4.29 pour obtenir l'énergie totale dans cette région du spectre.

$$(\{X_{train}\}_{mel})_u^l[\Lambda] = \sum_{k=0}^{N-1} b_{\Lambda}[k](\{X_{train}^{all}\}_{emph})_u^l[k] \quad (4.29)$$

Amplitude logarithmique

Puisque l'oreille humaine a une réponse logarithmique en fonction de l'amplitude des sons, le logarithme naturel de l'énergie obtenue doit être calculé au moyen de l'équation 4.30. La constante $\epsilon_{log} = 10^{-10}$ permet d'éviter de calculer le logarithme de zéro, advenant le cas peu probable que l'énergie soit nulle pour une bande du spectre. Un exemple des caractéristiques vocales ainsi obtenues est illustré à la figure 4.6. La position et l'amplitude des agglomérations d'énergie pour les phonèmes présentes dans le logarithme du spectre à la figure 4.3 sont également représentées dans cette figure, mais en utilisant des vecteurs à plus petite dimension.

$$(c_{train})_u^l[\Lambda] = \ln [(\{X_{train}\}_{mel})_u^l[\Lambda] + \epsilon_{log}] \quad (4.30)$$

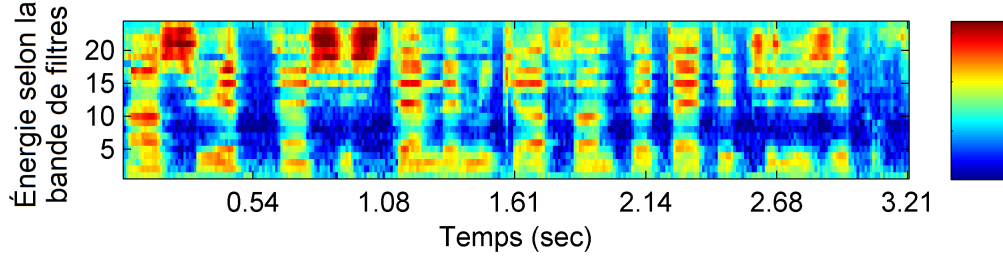


Figure 4.6 Exemple des caractéristiques vocales pour l'entraînement

Pour compléter cette étape du traitement, seules les trames d'indice l pour lesquelles l'activité vocale $(P_{vad})_u^l = 1$ sont conservées et sont renommées $(c_{train}^{speech})_u^l[k]$, pour un nouveau total de $(L_{train}^{speech})_u$ trames. Ce dernier tri permet d'obtenir les caractéristiques vocales actives pour l'entraînement dont un exemple est donné à la figure 4.7. On y voit que les trames inactives sont retirées de l'ensemble des caractéristiques.

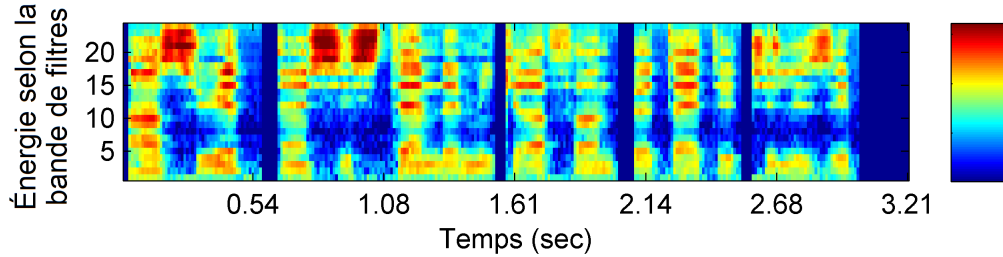


Figure 4.7 Exemple des caractéristiques vocales actives pour l'entraînement

Normalisation

Finalement, les caractéristiques vocales sont normalisées afin de réduire la distortion convolutive du canal (principalement reliée à la fonction de transfert du microphone utilisé pour l'enregistrement et au volume de la voix du locuteur). Dans le cas présent, la technique CMN est employée [16] et implantée par l'équation 4.31. Cette technique est appropriée car il n'y a pas de bruit additif provenant de l'environnement. Un exemple du résultat de cette normalisation est illustré à la figure 4.8. Dans cet exemple, les zones à faible énergie situées entre les dimensions 5 et 10 de la figure 4.7 possèdent un gain supérieur, ce qui permet de mieux représenter l'enveloppe du spectre.

$$(c_{train}^{cmn})_u^l[\Lambda] = (c_{train}^{speech})_u^l[\Lambda] - \frac{1}{(L_{train}^{speech})_u} \sum_{l=0}^{(L_{train}^{speech})_u-1} (c_{train}^{speech})_u^l[\Lambda] \quad (4.31)$$

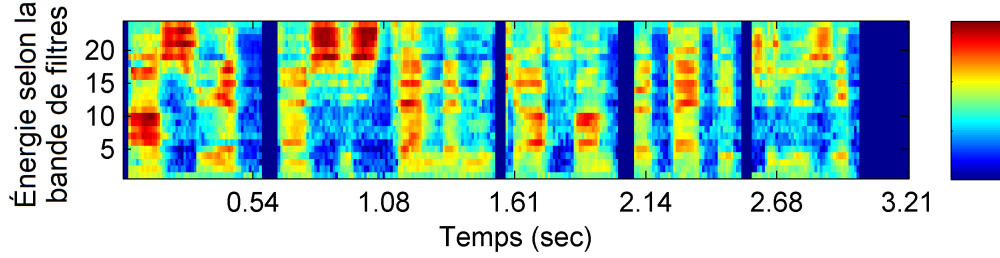
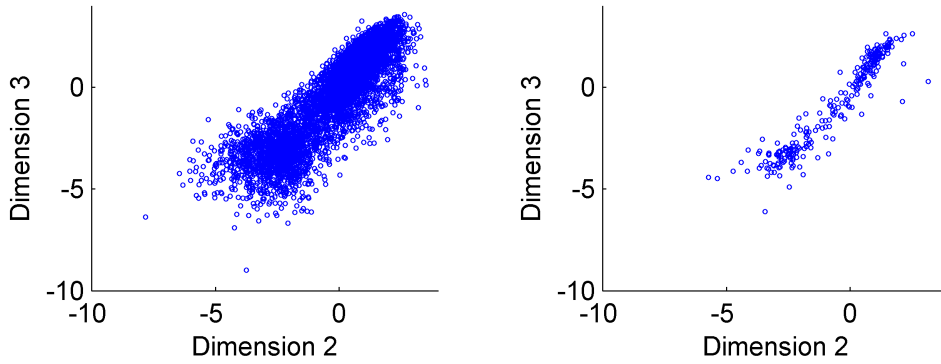


Figure 4.8 Exemple des caractéristiques vocales normalisées

4.1.3 Modèle

Un modèle est généré pour les caractéristiques obtenues à l'aide d'une technique de quantification vectorielle. Pour ce faire, l'algorithme k-moyennes est employé et accéléré grâce à l'inégalité triangulaire [11]. Un sous-ensemble de vecteurs κ_u^v est ainsi obtenu, pour un total de $V = 256$ vecteurs par modèle. Ce nombre est déterminé expérimentalement de façon à représenter adéquatement la distribution des caractéristiques tout en minimisant le nombre total de vecteurs. Chaque vecteur possède le même nombre de dimensions que les caractéristiques, c'est-à-dire $\Lambda_{max} = 24$ dimensions. L'expression $\kappa_u^v[\Lambda]$ désigne donc l'élément à la dimension Λ ($\Lambda = 0, \dots, \Lambda_{max} - 1$) du vecteur d'indice v ($v = 0, \dots, V - 1$) pour le modèle u . Un exemple de quantification vectorielle est illustré dans la figure 4.9 pour les deuxième et troisième dimensions.



(a) Énergie selon la bande de filtres des caractéristiques vocales normalisées ($(\mathbf{c}_{train}^{cmn})_u$) (b) Énergie selon la bande de filtres des centroïdes (κ_u)

Figure 4.9 Exemple de modélisation des caractéristiques vocales par quantification vectorielle

Habituellement, un GMM est employé pour ce genre d'application. L'avantage d'un GMM par rapport à une quantification vectorielle est qu'il inclut les deux premiers moments statistiques (moyenne et variance) plutôt que la moyenne uniquement comme pour la quantification vectorielle. Cependant, dans le cas présent, puisque les modèles sont mis

à jour de façon dynamique, une quantification vectorielle est plus appropriée. En effet, plusieurs techniques ont été développées pour mettre à jour la variance en fonction du bruit additif, mais en général l'information importante demeure dans la moyenne [41]. Par conséquent, la quantification vectorielle est utilisée pour l'application courante car elle offre de bonnes performances tout en étant une solution simple et efficace.

4.2 Reconnaissance de locuteurs avec WISS

La figure 4.10 illustre le traitement effectué par WISS pour la reconnaissance de locuteurs.

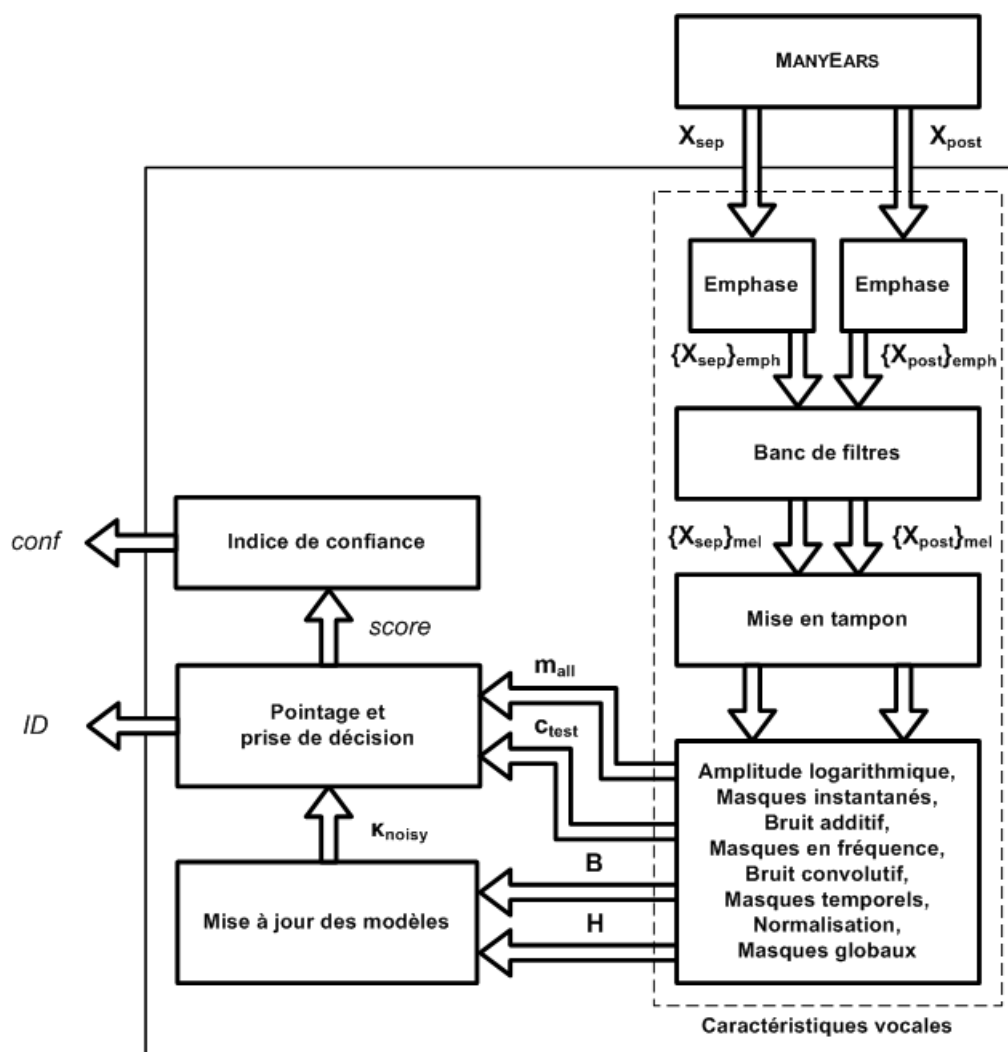


Figure 4.10 Schéma-bloc des modules pour la reconnaissance de locuteurs avec WISS

Dans le système mis en place, les segments de parole qui sont comparés aux modèles correspondant aux sources sonores séparées à partir du système ManyEars. Les signaux

post-filtrés sont également utilisés pour générer des masques. Les caractéristiques vocales pour la reconnaissance de locuteurs sont extraites du signal séparé pour chaque source sonore et à chacune des nouvelles trames du signal séparé. Cependant, chaque nouvelle caractéristique obtenue n'est pas immédiatement comparée avec un modèle de locuteurs (comme ce serait le cas dans un système purement temps réel avec un seul fil d'exécution). En effet, les caractéristiques sont d'abord mises en mémoire et sont par la suite normalisées lorsqu'un nombre suffisant de caractéristiques est obtenu. Cette contrainte provient du fait que certaines statistiques doivent être générées pour la normalisation, et que pour ce faire il faut donc une quantité minimale de caractéristiques.

Deux spectres sont utilisés pour générer les caractéristiques et les masques. Il s'agit des spectres des signaux séparés et post-filtrés qui sont calculés par les équations 4.32 et 4.33. La constante $(L_{test})_r$ représente le nombre de trames pour évaluer un segment.

$$(X_{test}^{sep})_r^l[k] = |(X_{sep})_s^l[k]|^2 \quad 0 \leq l < L_{test} \quad (4.32)$$

$$(X_{test}^{post})_r^l[k] = |(X_{post})_s^l[k]|^2 \quad 0 \leq l < L_{test} \quad (4.33)$$

Les opérations de fenêtrage et de transformée de Fourier discrète ne sont pas nécessaires à cette étape-ci de l'algorithme car elles ont déjà été effectuées précédemment par le système ManyEars. Cette simplification représente un avantage majeur car elle diminue la quantité de calculs et éliminent le problème de saturation qui peut survenir lorsque les signaux sont encodés en format 16 bits à la sortie du système ManyEars.

4.2.1 Caractéristiques vocales

Les caractéristiques obtenues sont identiques à celles présentées à la section 4.1.2. De plus, des masques instantanés, en fréquence, temporels et globaux sont générés. Ces derniers permettent d'estimer les bruit convolutifs et additifs, en plus d'être utilisés durant l'étape de pointage et prise de décision.

Emphase

Comme c'est le cas durant la phase d'entraînement, une emphase est appliquée au spectre du signal à traiter. Dans le cas présent, l'emphase est donc appliquée aux signaux séparés et post-filtrés par les équations 4.34 et 4.35. L'expression $H_{emph}[k]$ est identique à celle définie dans l'équation 4.22.

$$(\{X_{test}^{sep}\}_{emph})_r^l[k] = H_{emph}[k](X_{test}^{sep})_r^l[k] \quad (4.34)$$

$$(\{X_{test}^{post}\}_{emph})_r^l[k] = H_{emph}[k](X_{test}^{post})_r^l[k] \quad (4.35)$$

Un exemple du logarithme des signaux $(\{\mathbf{X}_{test}^{sep}\}_{emph})_r$ et $(\{\mathbf{X}_{test}^{post}\}_{emph})_r$ est illustré à la figure 4.11. Il est possible de voir à plusieurs temps différents des agglomérations d'énergie qui représentent différents phonèmes. Pour le signal séparé, on remarque la présence de bruit additif ce qui fait en sorte que le contraste entre les phonèmes est moins marqué. Dans le cas du signal post-filtré, ce phénomène est moins présent.

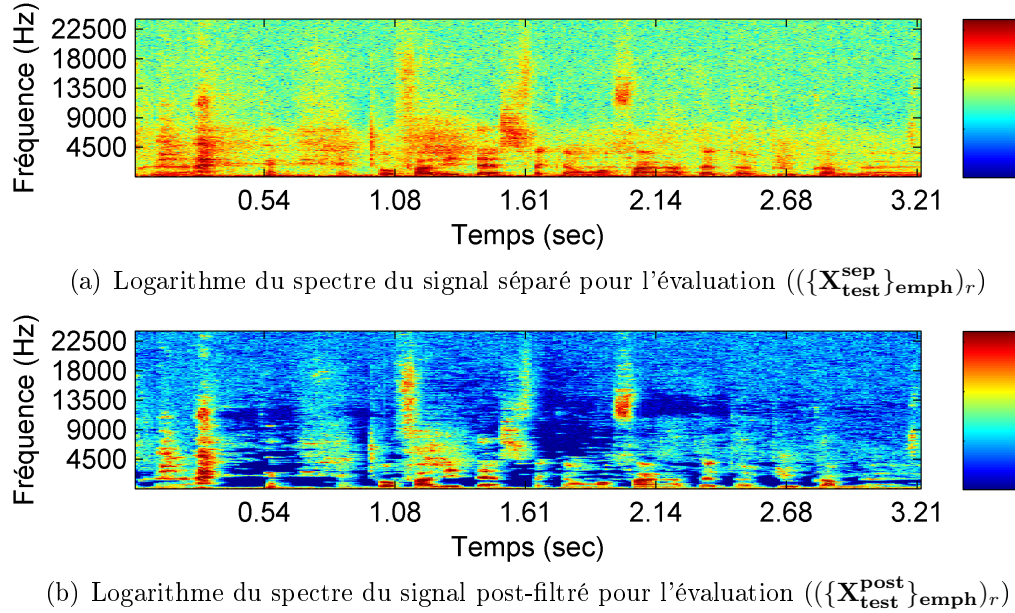


Figure 4.11 Exemple du logarithme du spectre des signaux pour l'évaluation

Banc de filtres

Par la suite, les bandes du spectre des deux signaux sont analysées à l'aide du banc de filtres défini par l'équation 4.28. Cette opération s'effectue par les équations 4.36 et 4.37.

$$(\{X_{test}^{sep}\}_{mel})_r^l[\Lambda] = \sum_{k=0}^{N-1} b_{\Lambda}[k] H_{emph}[k] (X_{test}^{sep})_r^l[k] \quad (4.36)$$

$$(\{X_{test}^{post}\}_{mel})_r^l[\Lambda] = \sum_{k=0}^{N-1} b_{\Lambda}[k] H_{emph}[k] (X_{test}^{post})_r^l[k] \quad (4.37)$$

Amplitude logarithmique

Tout comme pour l'entraînement, le logarithme de l'amplitude de l'énergie est calculé pour chaque bande avec l'équation 4.38. Cette opération n'est effectuée que sur le signal séparé, car c'est celui qui sert à la génération des caractéristiques vocales. Dans le cas présent, $\epsilon_{log} = 10^{-10}$. Contrairement à ce qui est fait par Valin et al. [46] pour la reconnaissance de la parole, le signal post-filtré n'est pas utilisé afin d'en extraire ses caractéristiques vocales. En effet, bien que le post-filtrage rehausse le rapport signal sur bruit, le spectre est déformé de manière variable dans le temps, ce qui supprime en bonne partie l'information propre à chaque locuteur.

$$(c_{test})_r^l[\Lambda] = \ln [(\{X_{test}^{sep}\}_{mel})_r^l[\Lambda] + \epsilon_{log}] \quad (4.38)$$

Un exemple de caractéristiques vocales en mode évaluation contaminés par le bruit additif est donné à la figure 4.12.

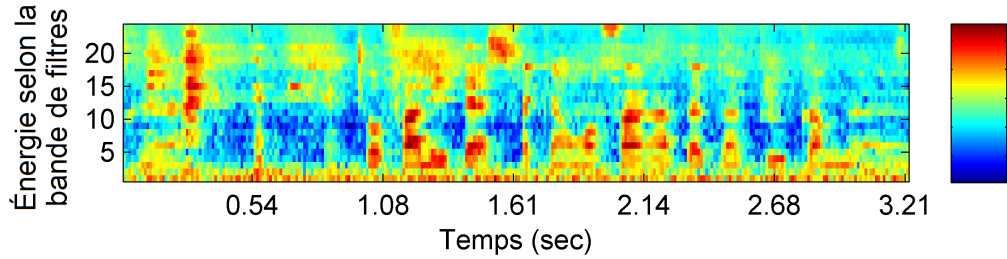


Figure 4.12 Exemple de caractéristiques vocales pour l'évaluation

Masques instantanés

Plusieurs masques sont générés à partir des signaux séparés et post-filtrés. Ces masques visent à ignorer certaines bandes du signal lorsque ces dernières sont corrompues par le bruit. Cette stratégie est particulièrement utile lorsque du bruit coloré est présent sur une partie du spectre seulement. L'usage des masques instantanés dépend du ratio entre les signaux post-filtré et séparé. Ce ratio est calculé par l'équation 4.39. Les masques binaires sont ainsi appliqués en fonction d'un seuil établi ($T_{inst} = 0.05$), tels que le décrit l'équation 4.40 [46].

$$(ratio_{inst})_r^l[\Lambda] = \frac{(\{X_{test}^{post}\}_{mel})_r^l[\Lambda]}{(\{X_{test}^{sep}\}_{mel})_r^l[\Lambda]} \quad (4.39)$$

$$(m_{inst})_r^l[\Lambda] = \begin{cases} 0 & (ratio_{inst})_r^l[\Lambda] < T_{inst} \\ 1 & (ratio_{inst})_r^l[\Lambda] \geq T_{inst} \end{cases} \quad (4.40)$$

Un exemple de masques instantanés est présenté à la figure 4.13. Les zones en blanc représentent une valeur unitaire pour le masque et les zones en noir représentent une valeur nulle. Il est observé que seules les agglomérations de la figure 4.12 qui possèdent une amplitude élevée se voient attribuer un masque non-nul.

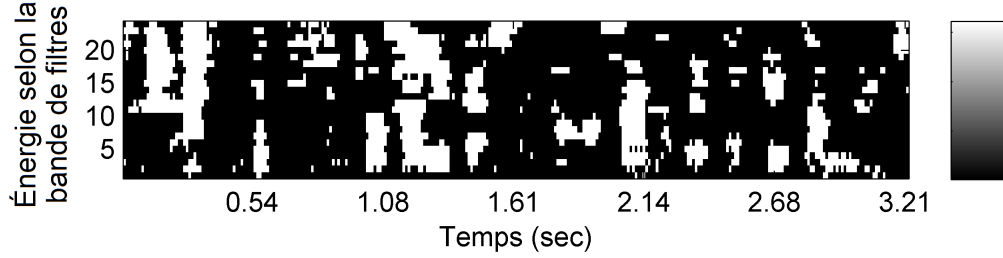


Figure 4.13 Exemple de masques instantanés

Bruit additif

Il est possible d'estimer le bruit additif $\hat{\mathbf{B}}_r$ durant les périodes de silence du locuteur. Les masques instantanés s'avèrent alors être un outil intéressant puisqu'ils indiquent les bandes qui sont dominées par la présence de bruit. Le bruit additif s'obtient à l'aide d'une moyenne pondérée calculée par l'équation 4.41. La variable $(L_{test})_r$ représente le nombre de trames pour le segment r .

$$\hat{B}_r[\Lambda] = \frac{\sum_{l=0}^{((L_{test})_r-1)} (c_{test})_r^l[\Lambda] (1 - (m_{inst})_r^l[\Lambda])}{\sum_{l=0}^{((L_{test})_r-1)} (1 - (m_{inst})_r^l[\Lambda])} \quad (4.41)$$

Un exemple de l'estimation du bruit additif est illustré à la figure 4.14. On constate que le bruit additif est surtout présent dans les basses fréquences. En effet, l'amplitude logarithmique du spectre est plus élevée pour les premières dimensions du vecteur.

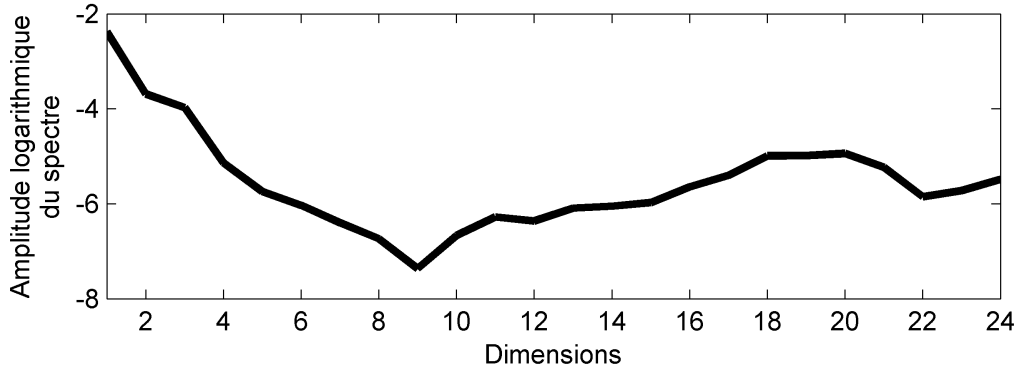


Figure 4.14 Exemple de l'estimation du bruit additif

Masques en fréquence

Les masques en fréquence sont utilisés pour déterminer la pertinence de chaque caractéristique vocale. En effet, il se peut parfois que la majorité des bandes soient corrompues par le bruit additif (dans les périodes de silence par exemple). Lorsque cela se produit, la caractéristique en question est considérée comme non pertinente et n'est pas utilisée pour la reconnaissance. Tel que décrit dans l'équation 4.42, la valeur 0 ou 1 de ce masque est obtenue en comparant la somme des composantes de chaque masque binaire à un seuil T_{freq} donné. Ce seuil a été expérimentalement fixé à 6. Ce type de masque est comparable à l'activité vocale illustrée précédemment à la figure 4.4.

$$(m_{freq})_r^l = \begin{cases} 0 & \left(\sum_{\Lambda=0}^{\Lambda_{max}-1} (m_{inst})_r^l[\Lambda] \right) < T_{freq} \\ 1 & \left(\sum_{\Lambda=0}^{\Lambda_{max}-1} (m_{inst})_r^l[\Lambda] \right) \geq T_{freq} \end{cases} \quad (4.42)$$

Un exemple de masques en fréquence est montré à la figure 4.15. Cette figure démontre que les trames de la figure 4.13 dont la somme de leurs masques est élevée sont considérées comme actives.

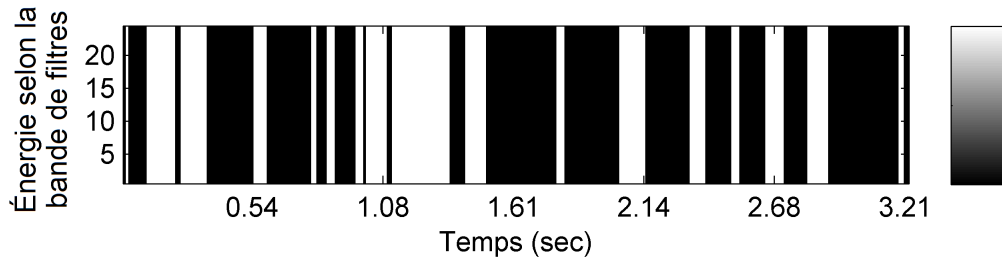


Figure 4.15 Exemple de masques en fréquence

Bruit convolutif

Le bruit convolutif est plus difficile à obtenir puisque la réverbération de la pièce dans laquelle un robot interagit avec des participants est généralement inconnue. Plusieurs techniques sont d'ailleurs proposées pour déterminer cette fonction de transfert à l'aide d'un processus d'optimisation [15]. Pour le cas présent, l'expérimentation a montré qu'il était possible d'obtenir une bonne estimation du bruit à partir des moyennes du signal bruité et du bruit additif.

Une estimation $(\hat{\mathbf{X}}_{\text{conv}})_r$ du signal bruité est obtenue en calculant la moyenne des caractéristiques avec les masques en fréquence. Seuls les segments de parole (et non ceux de silence) sont sélectionnés et représentés par $(c_{\text{test}}^{\text{speech}})_r^l[f]$. Il y a $(L_{\text{test}}^{\text{speech}})_r$ segments de parole qui correspondent aux masques en fréquence ayant une valeur non-nulle. Ce nombre est obtenu par l'équation 4.43.

$$(L_{\text{test}}^{\text{speech}})_r = \sum_{l=0}^{(L_{\text{test}})_r} (m_{\text{freq}})_r^l \quad (4.43)$$

La moyenne des segments de parole est obtenue par l'équation 4.44.

$$(\hat{X}_{\text{conv}})_r[\Lambda] = \frac{1}{(L_{\text{test}}^{\text{speech}})_r} \sum_{l=0}^{(L_{\text{test}}^{\text{speech}})_r-1} (c_{\text{test}}^{\text{speech}})_r^l[\Lambda] \quad (4.44)$$

La différence Ω entre le modèle κ_u et les caractéristiques $(\mathbf{c}_{\text{test}}^{\text{speech}})_r^l$ est représenté dans l'équation 4.45. La variable κ_u représente le modèle du locuteur dont l'indice est dénoté par la variable u . Chaque modèle est composé de plusieurs centroïdes, qui sont représentés par la variable κ_u^v dont l'indice est dénoté par la variable v . Bien que cette fonction ne soit pas la même que celle utilisée pour évaluer le pointage par rapport à chaque modèle, elle permet de simplifier les calculs et d'estimer de façon appropriée l'écart entre les modèles et les caractéristiques de test au cours du processus d'identification du bruit convolutif. Dans le cas présent, l'objectif, tel qu'énoncé par l'équation 4.46, est d'identifier le vecteur $\hat{\mathbf{H}}_r$ qui minimise la différence entre ces caractéristiques disponibles et l'ensemble des modèles des locuteurs.

$$\Omega = \sum_{l=0}^{(L_{\text{test}})_r-1} \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} \left\| \ln \left\{ \exp(\kappa_u^v + \mathbf{H}_r) + \exp(\hat{\mathbf{B}}_r) \right\} - (\mathbf{c}_{\text{test}}^{\text{speech}})_r^l \right\|^2 \quad (4.45)$$

$$\hat{\mathbf{H}}_r = \arg \min_{\mathbf{H}_r} [\Omega] \quad (4.46)$$

Au lieu de trouver directement une solution pour $\hat{\mathbf{H}}_r$ dans le but de minimiser la distortion, la variable intermédiaire $\hat{\mathbf{H}}_r^{u,v}$ est introduite. Il s'agit du bruit convolutif idéal pour minimiser la distortion pour le centroïde κ_u^v au vecteur v du modèle u . Cette variable est obtenue par l'équation 4.47.

$$\hat{\mathbf{H}}_r^{u,v} = \left[\sum_{u=0}^{U-1} \sum_{v=0}^{V-1} \left(\arg \min_{\mathbf{H}_r^{u,v}} \sum_{l=0}^{(L_{test})r-1} \left\| \ln \left\{ \exp(\kappa_u^v + \mathbf{H}_r^{u,v}) + \exp(\hat{\mathbf{B}}_r) \right\} - (\mathbf{c}_{\text{test}}^{\text{speech}})_r^l \right\|^2 \right) \right] \quad (4.47)$$

Il est possible de trouver un minimum grâce au gradient en trouvant la valeur de $\mathbf{H}_r^{u,v}$ qui correspond à la solution de l'équation 4.48, simplifiée à l'équation 4.49.

$$\nabla_{\mathbf{H}_r^{u,v}} \sum_{l=0}^{(L_{test})r-1} \left\| \ln \left\{ \exp(\kappa_u^v + \mathbf{H}_r^{u,v}) + \exp(\hat{\mathbf{B}}_r) \right\} - (\mathbf{c}_{\text{test}}^{\text{speech}})_r^l \right\|^2 = 0 \quad (4.48)$$

$$\sum_{l=0}^{(L_{test})r-1} \nabla_{\mathbf{H}_r^{u,v}} \left\| \ln \left\{ \exp(\kappa_u^v + \mathbf{H}_r^{u,v}) + \exp(\hat{\mathbf{B}}_r) \right\} - (\mathbf{c}_{\text{test}}^{\text{speech}})_r^l \right\|^2 = 0 \quad (4.49)$$

Le calcul du gradient mène à l'expression développée en 4.50. Cette expression se simplifie et l'équation 4.51 est ainsi obtenue.

$$\sum_{l=0}^{(L_{test})r-1} -2 \left(\ln \left\{ \exp(\kappa_u^v + \hat{\mathbf{H}}_r^{u,v}) + \exp(\hat{\mathbf{B}}_r) \right\} - (\mathbf{c}_{\text{test}}^{\text{speech}})_r^l \right) \left(\frac{\exp(\kappa_u^v + \hat{\mathbf{H}}_r^{u,v})}{\exp(\kappa_u^v + \hat{\mathbf{H}}_r^{u,v}) + \exp(\hat{\mathbf{B}}_r)} \right) = 0 \quad (4.50)$$

$$\ln \left\{ \exp(\kappa_u^v + \hat{\mathbf{H}}_r^{u,v}) + \exp(\hat{\mathbf{B}}_r) \right\} = \frac{1}{L_{test}^{\text{speech}}} \sum_{l=0}^{L_{test}-1} (\mathbf{c}_{\text{test}}^{\text{speech}})_r^l \quad (4.51)$$

En utilisant l'expression $(\hat{\mathbf{X}}_{\text{conv}})_r$ obtenue à l'équation 4.44, il est possible de trouver une expression simplifiée pour $\hat{\mathbf{H}}_r^{u,v}$ avec l'équation 4.52.

$$\hat{\mathbf{H}}_r^{u,v} = \ln \left\{ \exp((\hat{\mathbf{X}}_{\text{conv}})_r) - \exp(\hat{\mathbf{B}}_r) \right\} - \kappa_u^v \quad (4.52)$$

Des expressions optimales pour le bruit convolutif ($\mathbf{H}_r^{u,v}$) sont ainsi calculées pour chaque modèle u et vecteur v au sein de ces modèles. Bien entendu, une expression $\hat{\mathbf{H}}_r$ unique et commune à tous ces cas doit être obtenue. La stratégie pour y parvenir consiste à trouver une expression $\hat{\mathbf{H}}_r$ qui minimise la différence totale avec toutes les expressions $\mathbf{H}_r^{u,v}$ précédemment obtenues. Pour y parvenir, il s'agit de résoudre l'équation 4.53.

$$\hat{\mathbf{H}}_r = \arg \min_{\mathbf{H}_r} \left(\sum_{u=0}^{U-1} \sum_{v=0}^{V-1} \left\| \mathbf{H}_r - \hat{\mathbf{H}}_r^{u,v} \right\|^2 \right) \quad (4.53)$$

Le gradient est à nouveau utilisé pour résoudre ce problème d'optimisation, tel que formulé dans les équations 4.54 et 4.55.

$$\nabla_{\mathbf{H}_r^{u,v}} \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} \left\| \mathbf{H}_r - \hat{\mathbf{H}}_r^{u,v} \right\|^2 = 0 \quad (4.54)$$

$$\sum_{u=0}^{U-1} \sum_{v=0}^{V-1} \nabla_{\mathbf{H}_r^{u,v}} \left\| \mathbf{H}_r - \hat{\mathbf{H}}_r^{u,v} \right\|^2 = 0 \quad (4.55)$$

Le gradient est évalué en 4.56, pour finalement obtenir l'expression en 4.57.

$$\sum_{u=0}^{U-1} \sum_{v=0}^{V-1} 2 \left(\hat{\mathbf{H}}_r - \hat{\mathbf{H}}_r^{u,v} \right) = 0 \quad (4.56)$$

$$\hat{\mathbf{H}}_r = \frac{1}{UV} \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} \hat{\mathbf{H}}_r^{u,v} \quad (4.57)$$

Il est ensuite possible de substituer l'expression de $\mathbf{H}_r^{u,v}$ trouvée précédemment par l'équation 4.52 et ainsi obtenir l'équation 4.58.

$$\hat{\mathbf{H}}_r = \frac{1}{UV} \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} \left(\ln \left\{ \exp((\hat{\mathbf{X}}_{\text{conv}})_r) - \exp(\hat{\mathbf{B}}_r) \right\} - \kappa_u^v \right) \quad (4.58)$$

En simplifiant cette expression, l'équation 4.59 est obtenue. Cependant, puisqu'une normalisation a été initialement effectuée par l'équation 4.31 sur les caractéristiques vocales lors de l'entraînement des modèles, il est raisonnable d'affirmer par l'équation 4.60 que la moyenne des centroïdes est également nulle.

$$\hat{\mathbf{H}}_r = \ln \left\{ \exp((\hat{\mathbf{X}}_{\text{conv}})_r) - \exp(\hat{\mathbf{B}}_r) \right\} - \frac{1}{UV} \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} \kappa_u^v \quad (4.59)$$

$$\frac{1}{UV} \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} \kappa_u^v = 0 \quad (4.60)$$

L'expression finale pour l'estimation du bruit convolutif est donc obtenue à l'équation 4.61.

$$\hat{\mathbf{H}}_r = \ln \left\{ \exp((\hat{\mathbf{X}}_{\text{conv}})_r) - \exp(\hat{\mathbf{B}}_r) \right\} \quad (4.61)$$

Le bruit convolutif $\hat{\mathbf{H}}_r$ est donc estimé à partir de la différence (dans le domaine spectral linéaire) entre l'énergie du signal bruité et celle du bruit additif. Il est cependant possible que le bruit additif soit supérieur au bruit convolutif pour une ou plusieurs bandes. Advenant ce cas, une valeur nulle est attribuée au bruit convolutif de cette bande telle que montré dans l'équation 4.62. Pour ce faire, des masques temporels, présentés dans la section suivante, sont alors utilisés pour ignorer cette bande durant la comparaison avec les modèles de locuteurs.

$$\hat{H}_r[\Lambda] = \begin{cases} 0 & (\hat{X}_{\text{conv}})_r[\Lambda] \leq \hat{B}_r[\Lambda] \\ \ln \left\{ \exp \left((\hat{X}_{\text{conv}})_r[\Lambda] \right) - \exp \left(\hat{B}_r[\Lambda] \right) \right\} & (\hat{X}_{\text{conv}})_r[\Lambda] > \hat{B}_r[\Lambda] \end{cases} \quad (4.62)$$

Un exemple d'estimation de bruit convolutif est illustré à la figure 4.16. Dans cet exemple, la première bande possède une valeur nulle puisque le bruit additif est supérieur au bruit convolutif. Pour les autres bandes, le bruit convolutif est supérieur au bruit additif et est légèrement plus présent dans les basses fréquences.

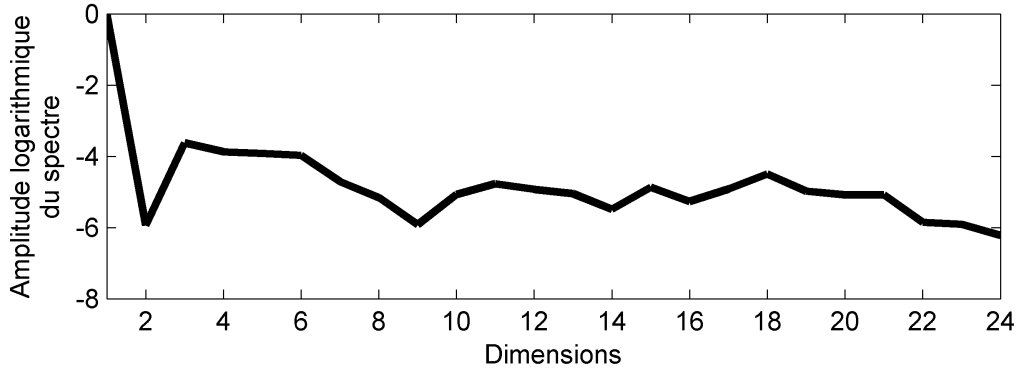


Figure 4.16 Exemple de l'estimation du bruit convolutif

Masques temporels

Le bruit convolutif $H_r[\Lambda]$ devrait excéder le bruit additif $B_r[\Lambda]$ pour chacune des bandes. Cependant, dans certaines situations comme en présence de bruit additif rose, il est possible que certaines bandes en basses fréquences soient complètement dominées par le bruit additif. Dans ce cas, ces bandes sont indésirables pour la reconnaissance et la normalisation spectrale, et doivent être négligées. Pour ce faire, un masque temporel est défini par les équations 4.63 et 4.64.

$$(ratio_{time})_r[\Lambda] = (\hat{X}_{conv})_r[\Lambda] - \hat{B}_r[\Lambda] \quad (4.63)$$

$$(m_{time})_r[\Lambda] = \begin{cases} 0 & (ratio_{time})_r[\Lambda] \leq 0 \\ 1 & (ratio_{time})_r[\Lambda] > 0 \end{cases} \quad (4.64)$$

Un exemple pour le masque temporel appliqué à la bande 0 est illustré dans la figure 4.17. Ceci est cohérent avec le fait que le bruit additif est supérieur au bruit convolutif, tel qu'illustré à la figure 4.16.

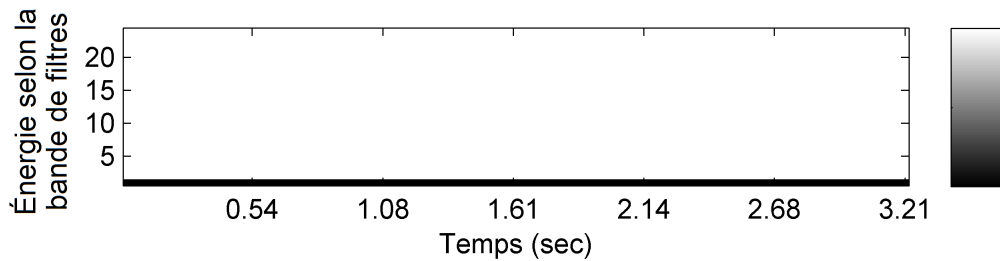


Figure 4.17 Exemple de masques temporels

Normalisation

La normalisation spectrale permet de conserver uniquement les variations d'amplitude d'une bande à l'autre et d'ignorer l'énergie totale du spectre. Cette procédure améliore la représentation de la forme du spectre, peu importe l'énergie instantanée de la trame en cours. Il s'agit donc de soustraire la moyenne de l'amplitude logarithmique des bandes à l'aide de l'équation 4.65. Dans cette équation, le calcul de la moyenne tient compte du masque temporel pour annuler la contribution des bandes corrompues par le bruit additif. Lorsque les coefficients MFCC sont utilisés dans le domaine cepstral, l'opération de normalisation s'effectue simplement en ignorant le premier terme de la transformée en cosinus discrète.

$$(c_{test}^{ac})_r^l[\Lambda] = (c_{test})_r^l[\Lambda] - \left(\sum_{\Lambda=0}^{\Lambda_{max}-1} (c_{test})_r^l[\Lambda] (m_{time})_r^l[\Lambda] \right) \quad (4.65)$$

Un exemple est illustré à la figure 4.18. On remarque que la variation d'énergie d'une trame à l'autre est moins prononcée comparativement à ce qui est illustré à la figure 4.12. Cette réduction de la variance découle directement de la soustraction de l'énergie instantanée à chaque trame.

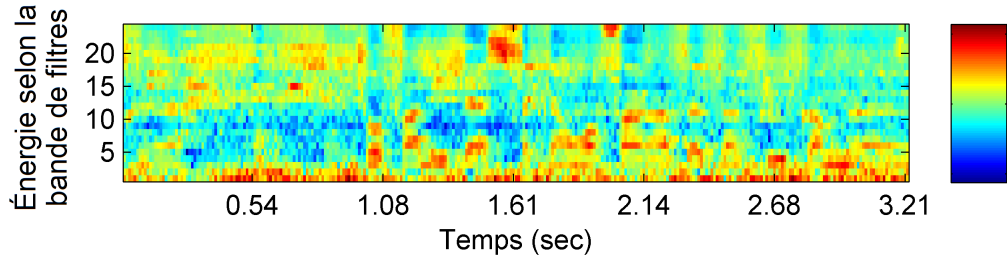


Figure 4.18 Exemple de caractéristiques vocales normalisées dans le domaine spectral

Masques globaux

Le masque global utilisé pour chaque caractéristique vocale est obtenu par l'équation 4.66 en multipliant les masques instantané, temporel et en fréquence. La figure 4.19 en illustre le mécanisme tandis que la figure 4.20 en donne un exemple.

$$(m_{all})_r^l[\Lambda] = (m_{freq})_r^l(m_{time})_r^l(m_{inst})_r^l[\Lambda] \quad (4.66)$$

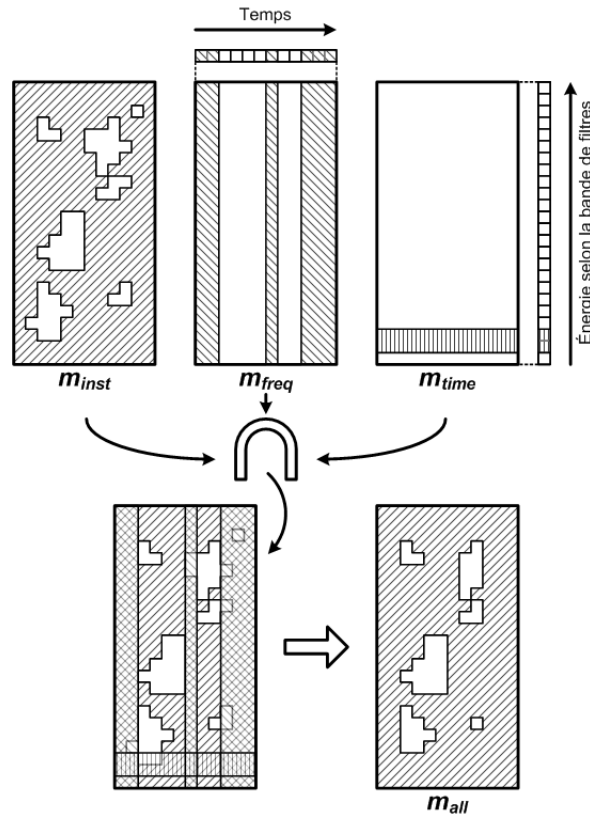


Figure 4.19 Combinaison des masques instantanés, temporels et en fréquence pour obtenir les masques globaux

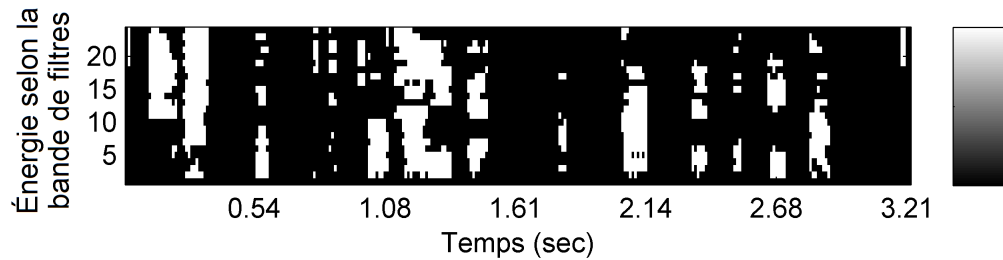


Figure 4.20 Exemple d'un masque global

4.2.2 Mise à jour des modèles

Pour obtenir des performances optimales, le modèle de chaque locuteur doit idéalement être entraîné dans le même environnement que celui dans lequel est faite la reconnaissance. Dans le cas d'un robot mobile, il est pratiquement impossible d'appliquer cette stratégie puisque l'environnement change constamment et ses propriétés ne peuvent être déterminées à l'avance. C'est pour cette raison que le modèle doit être entraîné à nouveau de façon dynamique dans les conditions de l'environnement actuel. Puisque des estimations des bruits convolutif et additif sont disponibles, il est possible de mettre à jour le modèle pour inclure ces nouvelles conditions. Il est cependant difficile d'entraîner le nouveau modèle à partir des caractéristiques vocales originales puisque ceci implique l'utilisation d'algorithmes tels que k-moyennes qui entraîne une charge importante de calculs. Le poids de ces calculs s'alourdit d'ailleurs en fonction du nombre de modèles à traiter. Pour cette raison, la technique PMC est utilisée afin de générer un nouveau modèle directement à partir des paramètres de l'ancien modèle entraîné dans des conditions idéales.

Durant l'entraînement, les caractéristiques vocales ne sont pas contaminées par le bruit additif et sont normalisées par rapport au gain du canal. Il est donc possible de contaminer ce modèle par les bruits convolutifs et additifs de l'environnement de test pour s'approcher d'un modèle qui aurait été caractérisé au départ dans cet environnement de test. Pour ce faire, le bruit convolutif est additionné dans le domaine spectral logarithmique et le bruit additif est additionné dans le domaine spectral linéaire. Cette opération est effectuée sur chaque vecteur du modèle et est décrite dans l'équation 4.67.

$$(\kappa_{noisy})_u^v[\Lambda] = \ln \left\{ \exp \left(\kappa_u^v[\Lambda] + \hat{H}_r[\Lambda] \right) + \exp \left(\hat{B}_r[\Lambda] \right) \right\} \quad (4.67)$$

Un exemple de modèle modifié est illustré à la figure 4.21. Il est intéressant de constater que le bruit additif affecte surtout les dimensions des caractéristiques qui possèdent de petites valeurs. Ceci a généralement pour effet de diminuer la variance car ces valeurs sont agglomérées lorsqu'elles sont à nouveau transformées dans le domaine spectral logarithmiques ($\ln(a + b) \approx \ln(a)$ si $b \ll a$).

Une fois cette opération complétée, le spectre des vecteurs du modèle peut être normalisé de la même manière que les caractéristiques vocales (équation 4.65) en utilisant l'équation 4.68.

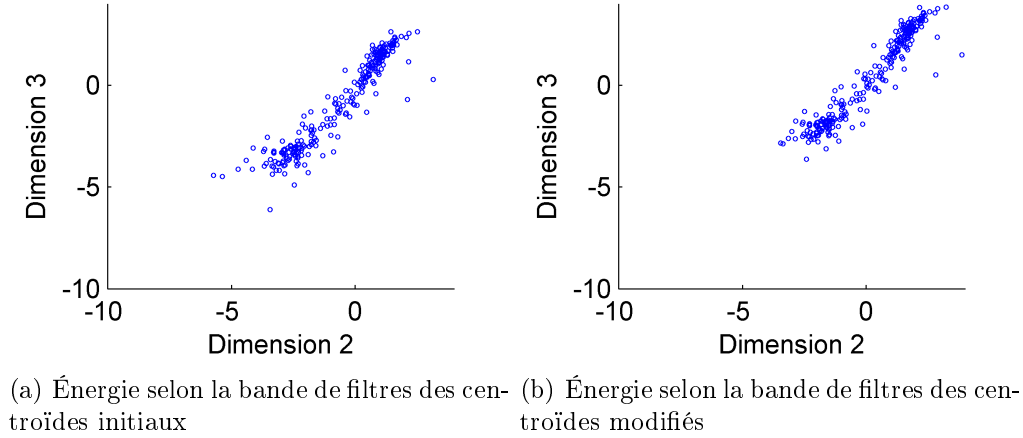


Figure 4.21 Modification dynamique des centroïdes

$$(\kappa_{ac})_u^v[\Lambda] = (\kappa_{noisy})_u^v[\Lambda] - \left(\sum_{\Lambda=0}^{\Lambda_{max}-1} (\kappa_{noisy})_u^v[\Lambda] (m_{time})_r[\Lambda] \right) \quad (4.68)$$

4.2.3 Pointage et prise de décision

Une fois que le modèle et les caractéristiques vocales du segment à identifier sont normalisés, il est possible d'effectuer une comparaison et de retourner un pointage. Ce dernier s'obtient par l'équation 4.69 en calculant la distance euclidienne entre la caractéristique en cours et le modèle sélectionné.

$$\text{dist}(\kappa, c, m) = \sqrt{\sum_{\Lambda=0}^{\Lambda_{max}-1} [m[\Lambda](\kappa[\Lambda] - c[\Lambda])^2]} \quad (4.69)$$

Il est à noter que c'est ici qu'est utilisé le masque binaire pour pondérer chaque bande. Le pointage de chaque modèle u par rapport au segment r est calculé en additionnant la distance minimale entre chaque caractéristique vocale $(c_{test}^{ac})_r^l$ et l'ensemble des vecteurs $(\kappa_{ac})_u^v$ du modèle u .

$$score_u^r = \sum_{l=0}^{L_{test}-1} \min_{v=0, \dots, (V-1)} \left\{ \text{dist}((\kappa_{ac})_u^v, (c_{test}^{ac})_r^l, (m_{all})_r^l) \right\} \quad (4.70)$$

L'identité du locuteur est finalement obtenue en sélectionnant le modèle qui offre le pointage minimal (c'est-à-dire que la différence entre les caractéristiques et le modèle est minimisée). Ceci est présenté dans l'équation 4.71.

$$(id_{exp})_r = \arg \min_u \{score_u^r\} \quad (4.71)$$

4.2.4 Indice de confiance

Il est également possible de considérer l'ajout d'un indice de confiance rattaché au locuteur identifié. Un tel indicateur agit comme complément à la décision par pointage et peut être utile entre autres dans un contexte de fusion, où la voix, la vidéo et d'autres informations biométriques sont présentes. Dans ce contexte, une identification avec un indice de confiance élevé aurait alors plus de poids parmi cet ensemble de données et affecterait plus directement le choix final de l'identité de l'individu.

L'indice de confiance retourné se situe entre 0 et 1. Une valeur de 0 représente une confiance nulle tandis qu'une valeur de 1 signifie une confiance absolue. Il est à noter qu'une confiance absolue n'implique pas forcément que l'identification soit valide, mais sous-entend que le système est certain d'avoir identifié correctement le locuteur parmi les modèles dont il disposait.

L'indice de confiance est basé sur l'écart entre les deux pointages les plus faibles représenté par la variable $\Delta score$ et s'évalue par l'équation 4.72. Les variables α_{score} et β_{score} ont été expérimentalement fixées à 0.02 et 0.01 respectivement.

$$conf_r = \frac{1}{1 + \exp [-(\Delta score - \alpha_{score})/\beta_{score}]} \quad (4.72)$$

La fonction sigmoïde ainsi obtenue est illustrée à la figure 4.22. Cette fonction est choisie car elle offre deux paliers de saturation avec une transition linéaire au centre. Une fonction échelon aurait pu être utilisée mais dans ce cas la transition aurait été trop abrupte entre une confiance nulle et absolue.

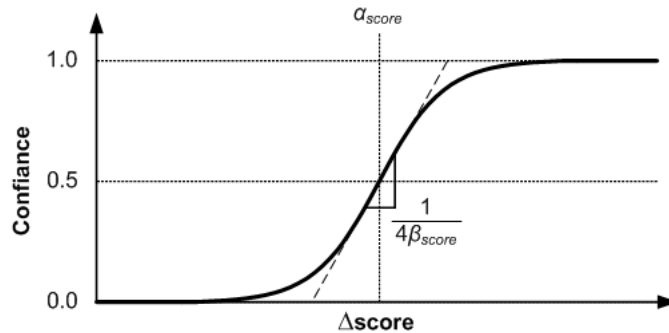


Figure 4.22 Courbe de fonction sigmoïde pour l'indice de confiance

4.3 Contributions

Cette nouvelle approche permet donc d'effectuer une reconnaissance de locuteurs dans un environnement bruyant. En effet, l'utilisation de caractéristiques MFCC dans le domaine spectral et de masques permet d'effectuer une estimation des bruits additif et convolutif. Ces estimations sont alors utilisées pour mettre à jour les modèles des locuteurs à l'aide de la technique PMC. De plus, l'utilisation des signaux propres au traitement réalisé par ManyEars permet d'ajouter une capacité de reconnaissance de locuteurs avec une faible augmentation en coût de calculs. Également, l'introduction d'un indice de confiance facilite le couplage de WISS avec d'autres systèmes d'identification biométriques dans le cadre de travaux futurs. L'analyse des performances de ce nouveau système est présentée au chapitre suivant.

CHAPITRE 5

ANALYSE DE PERFORMANCES

Le montage expérimental est le même que celui utilisé pour le système ManyEars [45]. Il est composé d'un ensemble de huit microphones, d'une carte d'acquisition de signaux analogiques et d'un ordinateur personnel de type PC. Les microphones sont disposés en cube et leurs positions respectives sont données au tableau 5.1. Cette configuration est sélectionnée car les microphones sont disposés d'une manière semblable sur le torse du robot (quatre microphones à l'avant et quatre à l'arrière). La carte audio, dotée d'un convertisseur analogique-numérique à 16 bits, effectue une acquisition synchrone sur huit canaux à une cadence de 48000 échantillons/sec par canal. Les données audio ainsi obtenues sont déposées en mémoire sur le disque dur de l'ordinateur, qui les analyse par la suite. La figure 5.1 illustre ce montage.

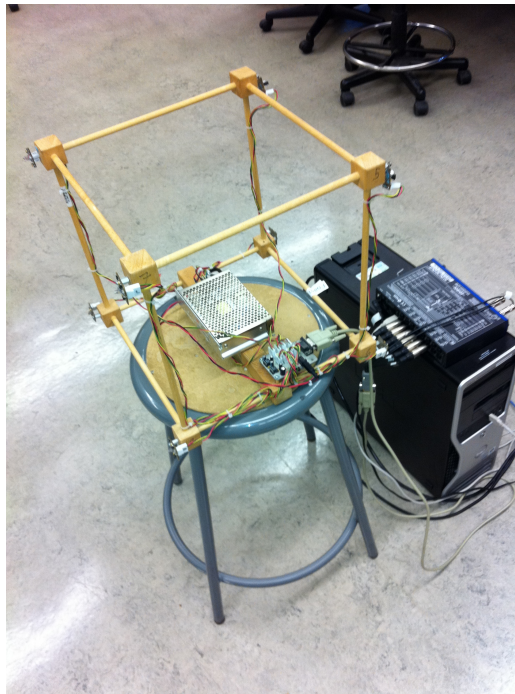


Figure 5.1 Microphones et système d'acquisition

Une série d'expériences ont été menées dans le local du laboratoire possédant une superficie de 10 mètres par 10 mètres et une hauteur de 2.5 mètres. Un système de ventilation ainsi que plusieurs appareils électriques (instruments, postes de travail, réfrigérateurs, etc) sont en opération continue dans cette pièce et génèrent du bruit audible. Le cube de

Tableau 5.1 Positions des microphones (en mètres)

m	$[(z_{mic})_m]_x$	$[(z_{mic})_m]_y$	$[(z_{mic})_m]_z$
0	+0.16	+0.16	+0.16
1	+0.16	+0.16	-0.16
2	+0.16	-0.16	+0.16
3	+0.16	-0.16	-0.16
4	-0.16	+0.16	+0.16
5	-0.16	+0.16	-0.16
6	-0.16	-0.16	+0.16
7	-0.16	-0.16	-0.16

microphones est installé au centre de la pièce à une hauteur d'environ soixante centimètres. Cette hauteur est réaliste en ce qui concerne l'emplacement des microphones sur un robot.

Les performances de WISS sont mesurées selon deux indices. Le premier est l'indice non-pondéré ($rate_{noWeight}$), qui consiste à calculer le taux d'identification selon le nombre de bonnes identifications ($(match_{good}^{noWeight})_r$) et de mauvaises identifications ($(match_{bad}^{noWeight})_r$) tel que décrit dans les équations 5.1, 5.2 et 5.3.

$$(match_{good}^{noWeight})_r = \begin{cases} 0 & (id_{exp})_r \neq (id_{theo})_r \\ 1 & (id_{exp})_r = (id_{theo})_r \end{cases} \quad (5.1)$$

$$(match_{bad}^{noWeight})_r = \begin{cases} 1 & (id_{exp})_r \neq (id_{theo})_r \\ 0 & (id_{exp})_r = (id_{theo})_r \end{cases} \quad (5.2)$$

$$rate_{noWeight} = \frac{\sum_{r=0}^{R-1} (match_{good}^{noWeight})_r}{\sum_{r=0}^{R-1} (match_{good}^{noWeight})_r + \sum_{r=0}^{R-1} (match_{bad}^{noWeight})_r} \quad (5.3)$$

Le second indice est dit pondéré ($rate_{weight}$), car le taux d'identification dépend de l'indice de confiance ($conf_r$) envers les bonnes ($(match_{good}^{weight})_r$) et les mauvaises identifications ($(match_{bad}^{weight})_r$). Le calcul de cet indice est réalisé par les équations 5.4, 5.5 et 5.6.

$$(match_{good}^{weight})_r = \begin{cases} 0 & (id_{exp})_r \neq (id_{theo})_r \\ conf_r & (id_{exp})_r = (id_{theo})_r \end{cases} \quad (5.4)$$

$$(match_{bad}^{weight})_r = \begin{cases} conf_r & (id_{exp})_r \neq (id_{theo})_r \\ 0 & (id_{exp})_r = (id_{theo})_r \end{cases} \quad (5.5)$$

$$rate_{weight} = \frac{\sum_{r=0}^{R-1} (match_{good}^{weight})_r}{\sum_{r=0}^{R-1} (match_{good}^{weight})_r + \sum_{r=0}^{R-1} (match_{bad}^{weight})_r} \quad (5.6)$$

Trois scénarios de tests ont été élaborés pour évaluer les capacités de WISS :

1. Caractérisation du système à partir d'une banque de données de vingt locuteurs
2. Interaction statique dans des conditions plus réalistes où un seul locuteur parle à la fois
3. Interaction dynamique avec plusieurs locuteurs qui participent à une discussion naturelle

5.1 Caractérisation du système

Pour ce scénario de tests, des segments de parole sont diffusés dans un haut-parleur placé à une position donnée par rapport au cube de microphones. Un corpus de données de vingt locuteurs composé de onze voix de femmes et neuf voix d'hommes est utilisé. Pour chaque locuteur, un modèle est entraîné directement à partir d'un segment de voix de soixante secondes non-bruité en provenance du corpus de données. Par la suite, six segments de parole de dix secondes sont diffusés dans le haut-parleur et captés par le cube de microphones. Pour chacun des segments, cinq niveaux de volume différents (numérotés de 1 à 5) sont sélectionnés. Ceci a pour but de vérifier que le système peut composer avec un volume de locuteur variable en présence de bruit additif (par exemple les ventilateurs) à volume constant. Bien entendu, les segments de parole utilisés pour les tests sont différents de ceux utilisés pour l'entraînement des modèles.

Dans le but de vérifier l'effet du formateur de faisceaux sur le SNR, trois signaux sont évalués. Le premier, $(X_{speaker}^{signalMean})_s^l[k]$, provient de l'énergie de la somme de tous les signaux des microphones (5.7), le second, $(X_{speaker}^{powerMean})_s^l[k]$, de la somme de l'énergie de tous les microphones (5.8) et le troisième, $(X_{speaker}^{beamformer})_s^l[k]$, de la sortie de la séparation du formateur de faisceaux (5.9). Le signal d'un seul microphone n'est pas utilisé car ces microphones ne sont pas parfaitement omnidirectionnels, ce qui affecte la qualité sonore selon la position de la source. Le SNR est estimé pour chacun de ces trois signaux au cours des périodes de silence et de parole : *signalMean* représente une interférence aléatoire

(les signaux des microphones sont simplement additionnés sans considérer la position de la source); *powerMean* représente la puissance moyenne sans interférence, et c'est pour cette raison que les puissances et non les signaux initiaux sont additionnées; *beamformer* représente finalement le résultat d'une interférence constructive. L'équation mathématique du SNR est donnée en 5.10 (l'expression *mode* est remplacée par *signalMean*, *powerMean* et *beamformer*).

$$(X_{speaker}^{signalMean})_s^l[k] = \left| \sum_{m=0}^{M-1} (X_{mic})_m^l[k] \right|^2 \quad (5.7)$$

$$(X_{speaker}^{powerMean})_s^l[k] = \sum_{m=0}^{M-1} |(X_{mic})_m^l[k]|^2 \quad (5.8)$$

$$(X_{speaker}^{beamformer})_s^l[k] = |(X_{sep})_s^l[k]|^2 \quad (5.9)$$

$$SNR \approx \frac{(L_{noise})_s \sum_{l=0}^{(L_{speech})_s} \sum_{k=0}^{N-1} |\{(X_{speaker}^{mode})_s^l[k]\}_{speech}|^2}{(L_{speech})_s \sum_{l=0}^{(L_{noise})_s} \sum_{k=0}^{N-1} |\{(X_{speaker}^{mode})_s^l[k]\}_{noise}|^2} \quad (5.10)$$

Les performances sont évaluées en fonction de l'angle de la source par rapport au robot dans la section 5.1.1 et selon la distance entre la source et le robot dans la section 5.1.2.

5.1.1 Performances selon la position angulaire

Cette expérience a pour but de valider le bon fonctionnement du système pour toutes orientations du locuteur par rapport au robot. Le cube de microphones est placé au centre d'un cercle de rayon de 1.5 mètres, sur lequel sont sélectionnées huit positions pour le haut-parleur. Celles-ci sont séparées l'une de l'autre par un écart angulaire de quarante-cinq degrés, tel que montré à la figure 5.2. Pour chacune des positions, le haut-parleur est installé à une hauteur d'environ soixante centimètres.

Les SNRs sont présentés au tableau 5.2. On constate d'abord que les SNRs sont similaires peu importe la position angulaire, ce qui signifie que le système ne dépend pas de l'orientation du locuteur. Il est intéressant de constater que le SNR du signal *beamformer* est, tel qu'anticipé, supérieur à celui du signal *signalMean*. Par contre, on constate qu'il est similaire à celui du *powerMean*, ce qui peut paraître surprenant. En effet, il est normal d'imaginer *a priori* que l'interférence constructive dans la direction de la source sonore

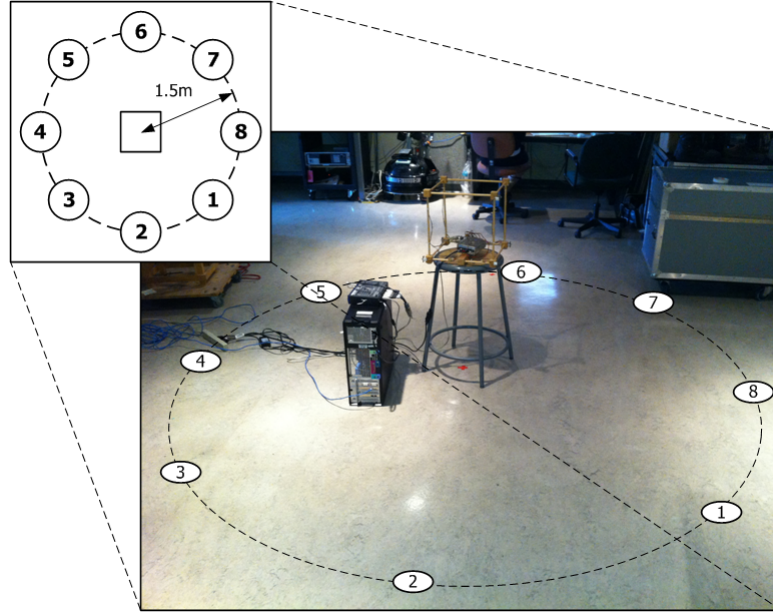


Figure 5.2 Positions angulaires

offre un meilleur SNR qu'une interférence non constructive. Il faut cependant réaliser que le formateur de faisceaux utilisé par le système ManyEars vise avant tout à réduire le bruit directionnel (par exemple un autre locuteur qui parle simultanément) et non le bruit omnidirectionnel ambiant (par exemple des ventilateurs qui sont assez éloignés pour que le bruit soit diffusé de façon aléatoire dans la pièce). Il existe des méthodes pour optimiser un formateur de faisceaux afin de réduire au maximum le bruit additif. Cependant, dans le présent contexte, on considère que le formateur de faisceaux du système ManyEars est plus adéquat car il offre d'excellentes performances pour la reconnaissance de parole à plusieurs locuteurs tout en offrant des SNRs similaires au cas sans interférence (*powerMean*). Les SNRs de *powerMean* sont parfois supérieurs à *beamformer*, entre autres pour les positions 3, 6 et 8. Puisque ces deux signaux sont similaires, la variation de la réverbération et le niveau de bruit ambiant selon la position peut influencer le SNR et explique pourquoi le signal *powerMean* offre parfois un SNR légèrement supérieur à celui de *beamformer*.

Une première reconnaissance est effectuée sans l'utilisation des masques durant l'étape du pointage, comme l'exprime l'équation 4.69. Ceci revient donc à appliquer un masque de valeur de 1 pour toutes les dimensions. Les performances obtenues avec les indices non-pondérées ($rate_{noWeight}$) et pondérées ($rate_{weight}$) sont présentées aux figures 5.3 et 5.4. Par la suite, les masques sont utilisés et les performances obtenues avec les indices non-pondérées ($rate_{noWeight}$) et pondérées ($rate_{weight}$) sont présentées aux figures 5.5 et 5.6.

Tableau 5.2 SNR selon la position angulaire

Position		1	2	3	4	5	6	7	8
signalMean	Niveau 1	11.41	13.31	12.50	12.04	11.39	12.65	12.48	11.48
	Niveau 2	7.20	8.78	8.18	7.72	7.20	8.35	8.14	7.19
	Niveau 3	3.93	5.13	4.60	4.19	3.77	4.70	4.60	4.00
	Niveau 4	1.76	2.45	2.17	1.99	1.75	2.23	2.15	1.75
	Niveau 5	0.81	1.15	0.72	0.78	0.74	1.05	0.95	0.58
powerMean	Niveau 1	16.87	16.89	16.92	14.80	15.01	15.97	17.13	16.91
	Niveau 2	12.13	12.08	12.26	10.22	10.45	11.39	12.43	12.11
	Niveau 3	7.92	7.90	7.96	6.16	6.30	7.15	8.15	8.00
	Niveau 4	4.35	4.35	4.42	3.23	3.33	3.87	4.54	4.39
	Niveau 5	2.14	2.14	1.90	1.39	1.51	1.92	2.22	1.94
beamformer	Niveau 1	16.92	17.33	16.56	15.56	15.31	15.75	17.29	16.59
	Niveau 2	12.20	12.57	11.86	10.91	10.77	11.21	12.54	11.84
	Niveau 3	7.94	8.23	7.72	6.68	6.60	6.99	8.24	7.75
	Niveau 4	4.41	4.61	4.27	3.58	3.52	3.66	4.60	4.18
	Niveau 5	2.07	2.25	1.82	1.50	1.61	1.63	2.24	1.69

Ces graphiques illustrent que les performances restent essentiellement les mêmes pour les différentes positions angulaires. Ceci indique clairement que le système est en mesure de reconnaître les locuteurs peu importe leur orientation par rapport au robot. Cette caractéristique est importante car elle permet au robot d’interagir sans avoir à continuellement s’orienter pour faire face à la personne qui lui parle si cette dernière se déplace de temps à autre. De plus, les performances diminuent lorsque le SNR est moins élevé. Le taux de reconnaissance est obtenu à l’aide de la moyenne des indices pondérés des positions angulaires 1 à 8 avec le formateur de faisceaux avec masques. Il se situe à 95.6% au niveau 1 (le SNR moyen est de 16 dB) et diminue à 84.3% au niveau 5 (le SNR moyen est de 2 dB). On constate d’ailleurs une dégradation importante des performances pour les niveaux 4 et 5, qui s’explique par des faibles SNRs d’approximativement 4dB et 2dB respectivement. Les performances sont meilleures pour les figures 5.5 et 5.6 comparativement aux figures 5.3 et 5.4, ce qui démontre que l’utilisation de masques offre des performances supérieures par rapport à un système sans masque, en particulier dans des cas où le SNR est faible. Cette caractéristique est particulièrement intéressante car les masques sont également utilisés dans un contexte de reconnaissance de la parole, ce qui permet deux applications différentes avec un prétraitement commun.

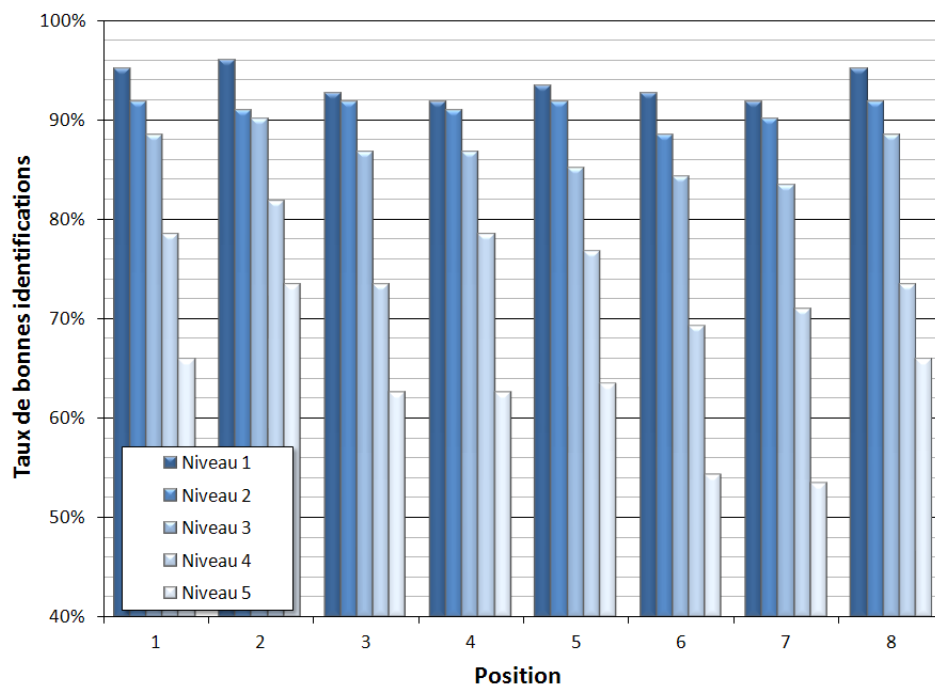


Figure 5.3 Performances avec les indices non-pondérés pour les huit positions angulaires avec le formateur de faisceaux sans masques

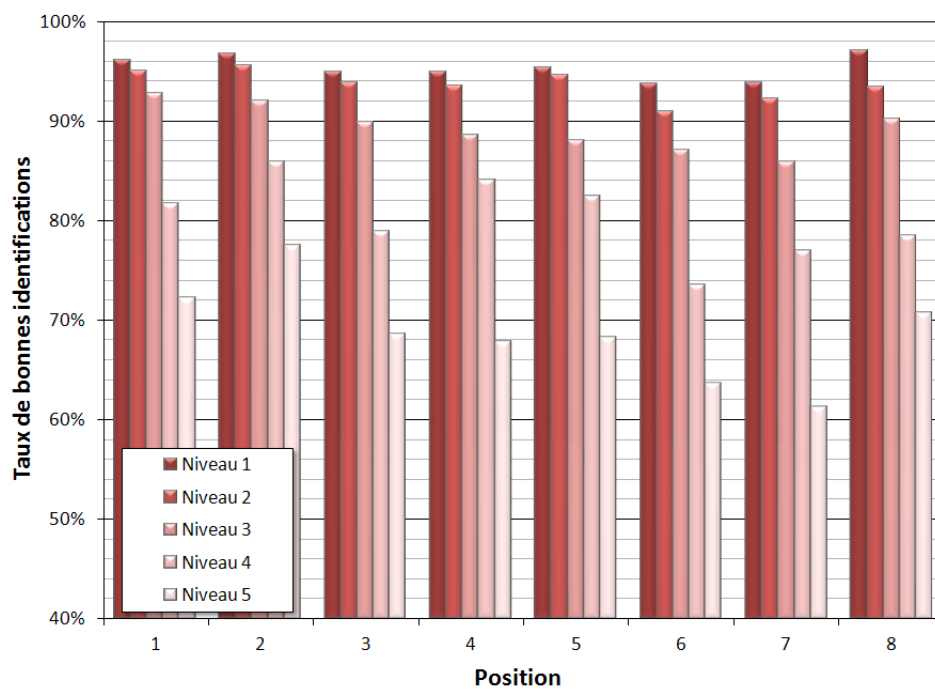


Figure 5.4 Performances avec les indices pondérés pour les huit positions angulaires avec le formateur de faisceaux sans masques

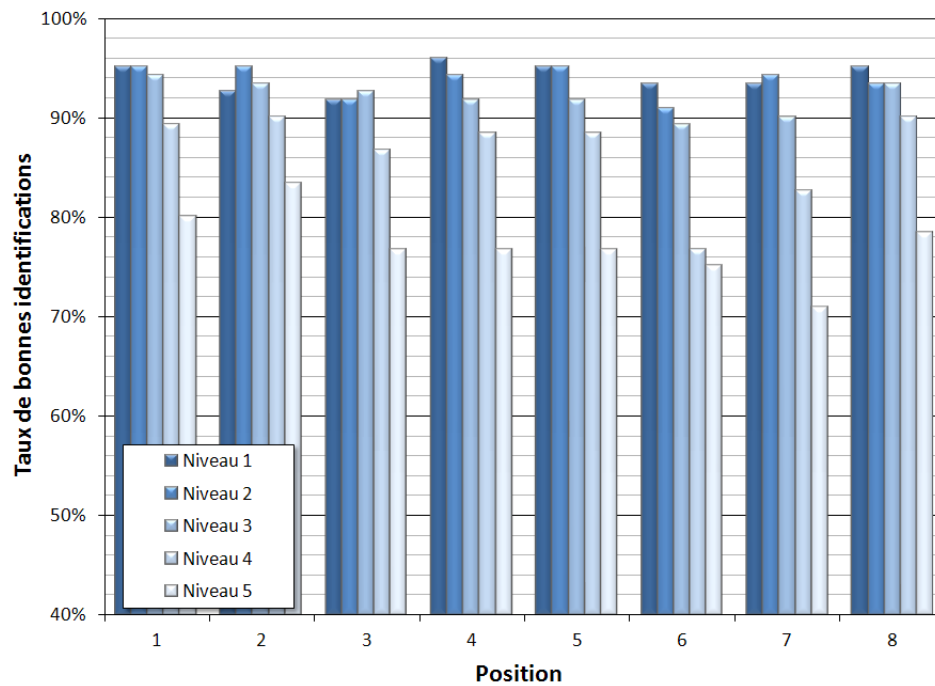


Figure 5.5 Performances avec les indices non-pondérés pour les huit positions angulaires avec le formateur de faisceaux avec masques

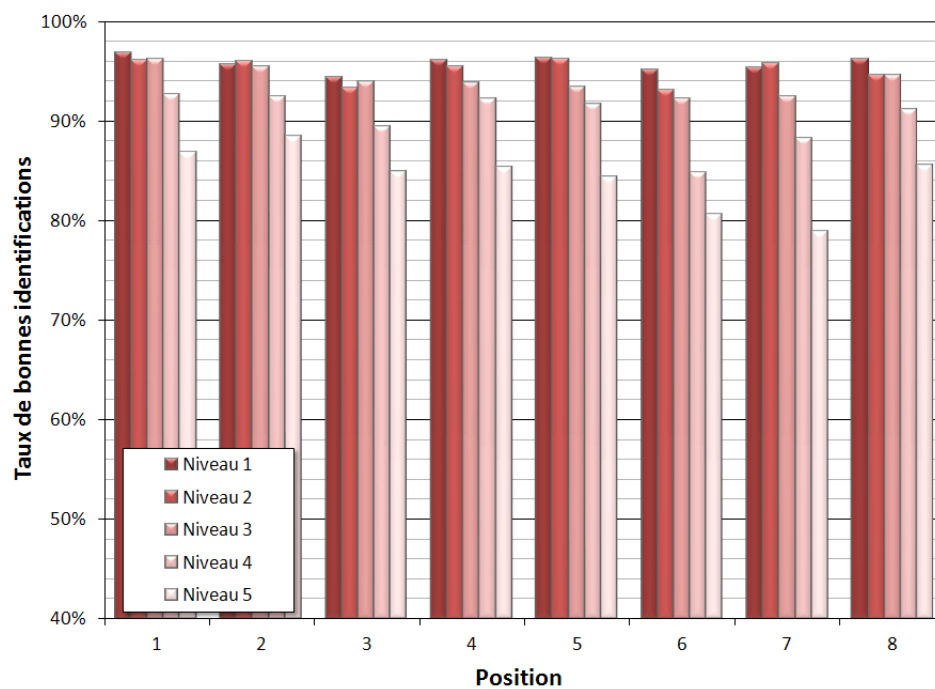


Figure 5.6 Performances avec les indices pondérés pour les positions angulaires avec le formateur de faisceaux avec masques

5.1.2 Performances selon la position radiale

Cette expérience permet d'observer la dégradation anticipée des performances lorsque le locuteur s'éloigne du robot. Trois positions différentes alignées à des distances de 1.5 mètres, 2.5 mètres et 3.5 mètres du cube de microphones sont sélectionnées. Ces positions sont présentées à la figure 5.7.

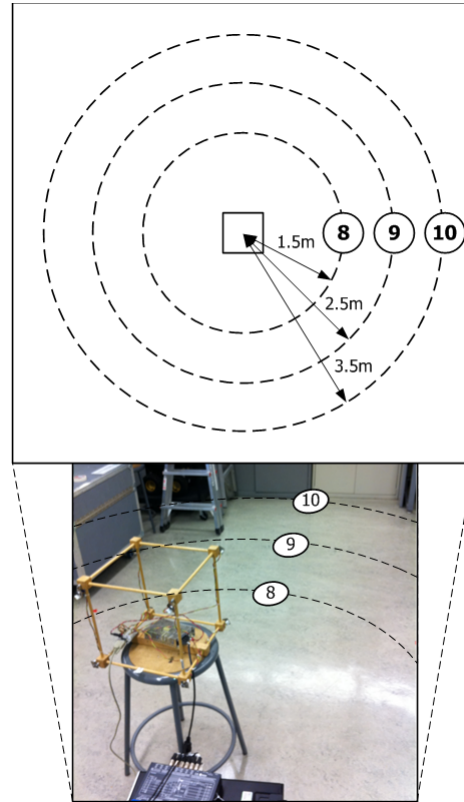


Figure 5.7 Positions radiales

Les SNRs sont à nouveau obtenus grâce à l'équation 5.10 et sont affichés au tableau 5.3. Comme c'est le cas à la section 5.1.1, le SNR est mesuré pour trois cas différents (*signalMean*, *powerMean* et *beamformer*). Encore une fois, le SNR du signal *beamformer* est supérieur à celui du signal *signalMean* et est similaire à celui du signal *powerMean* (qui demeure toutefois toujours supérieur). Pour les mêmes raisons que celles évoquées dans la section 5.1.2, le signal *beamformer* est utilisé afin d'en extraire les caractéristiques vocales.

Comme à la section 5.1.1, les performances des indices non-pondérées et pondérées sans l'utilisation des masques sont présentées aux figures 5.8 et 5.9. Par la suite, les performances sont évaluées avec l'utilisation des masques et les résultats des indices non-pondérés et pondérés sont montrés aux figures 5.10 et 5.11.

Tableau 5.3 SNR selon la position radiale

Position		8	9	10
signalMean	Niveau 1	11.48	8.60	7.54
	Niveau 2	7.19	4.96	4.10
	Niveau 3	4.00	2.28	1.84
	Niveau 4	1.75	0.98	0.77
	Niveau 5	0.58	0.33	0.27
powerMean	Niveau 1	16.92	13.42	11.64
	Niveau 2	12.11	9.03	7.39
	Niveau 3	8.00	5.12	3.93
	Niveau 4	4.39	2.57	1.84
	Niveau 5	1.94	1.01	0.71
beamformer	Niveau 1	16.59	12.48	10.22
	Niveau 2	11.84	8.13	6.14
	Niveau 3	7.75	4.50	3.11
	Niveau 4	4.18	2.18	1.30
	Niveau 5	1.69	0.83	0.42

Le tableau 5.3 soutient aussi que le SNR diminue lorsque la distance augmente. Ceci implique donc que la présence du bruit additif est plus significative lorsque le locuteur se situe à plusieurs mètres du robot. Le SNR du niveau 5 à la position 10 (distance de 3.5 mètres) est particulièrement faible puisqu'il se rapproche du 0 dB, ce qui signifie que le bruit est presque aussi présent en terme d'énergie que le signal du locuteur.

Les résultats présentés aux figures 5.8, 5.9, 5.10 et 5.11 suggèrent que les performances diminuent lorsque le locuteur s'éloigne des microphones. Il est intéressant de noter aux figures 5.8 et 5.10 que les performances semblent meilleures lorsque le locuteur se situe à la position 8 (distance de 2.5 mètres) par rapport à la position 7 (distance de 1.5 mètres). Cette légère différence s'explique par le fait que l'utilisation d'un indice non-pondéré entraîne un changement abrupte dans les performances lorsque deux modèles obtiennent un pointage similaire pour un même locuteur. Dans le cas actuel, le changement de position diminue le SNR, ce qui introduit une plus grande incertitude au niveau du pointage, et mène par conséquent à des décisions imprécises.

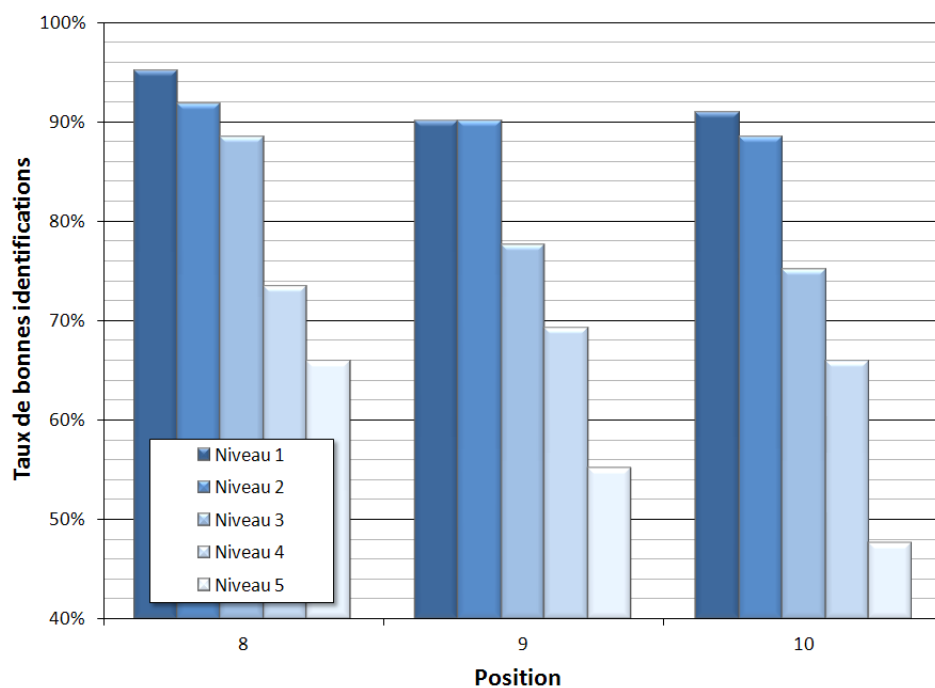


Figure 5.8 Performances avec les indices non-pondérés pour les positions radiales avec le formateur de faisceaux sans masques

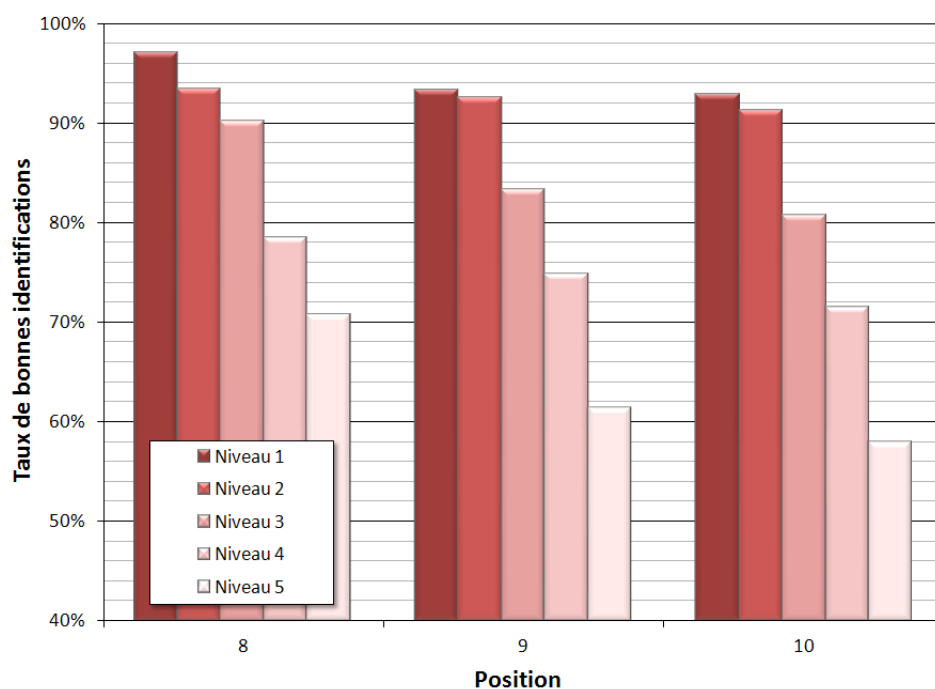


Figure 5.9 Performances avec les indices pondérés pour les positions radiales avec le formateur de faisceaux sans masques

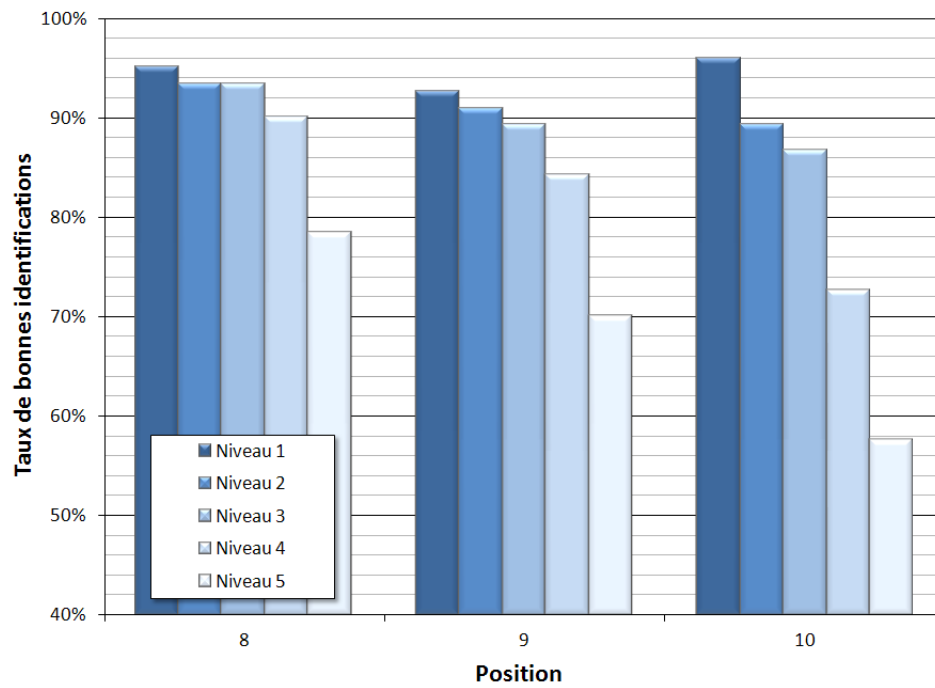


Figure 5.10 Performances avec les indices non-pondérés pour les positions radiales avec le formateur de faisceaux avec masques

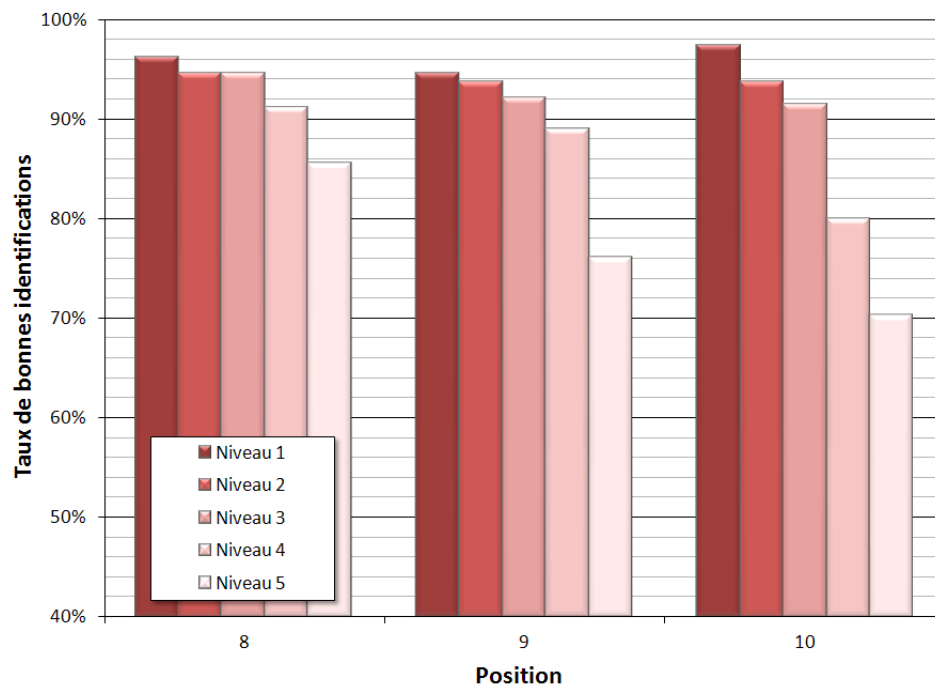


Figure 5.11 Performances avec les indices pondérés pour les positions radiales avec le formateur de faisceaux avec masques

5.1.3 Performances sur l'ensemble des positions

Pour chaque méthode proposée, les moyennes des performances pour l'ensemble des positions angulaires et radiales décrites aux sections 5.1.1 et 5.1.2 sont présentées à la figure 5.12. Les résultats suggèrent que l'utilisation de masques et d'un indice pondéré contribuent à améliorer les performances, en particulier lorsque le niveau de bruit est élevé.

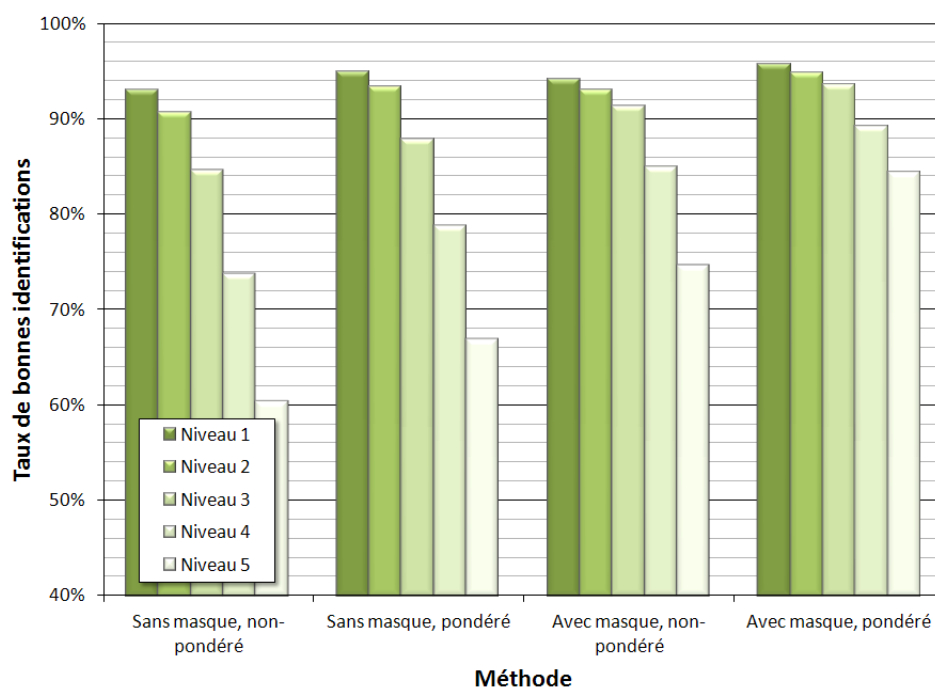


Figure 5.12 Moyennes des taux de bonnes identifications selon la méthode sélectionnée

5.2 Interaction statique

Cette expérience a pour but de vérifier les performances lorsqu'il s'agit d'une interaction entre un seul locuteur et le robot dans un environnement bruité. Pour entraîner les modèles, quatre locuteurs mâles récitent un texte aléatoire d'une durée d'environ soixante secondes en étant à une distance d'environ vingt centimètres d'un des huit microphones du cube. Bien que cet entraînement soit effectué dans un environnement bruité (ventilateurs, appareils électroniques, etc.), la proximité du locuteur par rapport au microphone fait en sorte que le bruit additif est négligeable, ce qui est nécessaire pour l'étape de l'entraînement. Un modèle est par la suite généré pour chacun des quatre locuteurs (identifié par les lettres A, B, C et D). La nouvelle banque de données est composée de ces quatre locuteurs seulement. Le nombre de locuteurs dans la banque de données est moins important que pour l'expérience de la section 5.1 mais correspond au nombre d'individus avec lesquels un robot pourrait interagir dans un domicile. Chaque locuteur est ensuite positionné à une distance de 1.5 mètres du cube et effectue la lecture d'une dizaine de phrases aléatoires d'une durée entre cinq et dix secondes dans ce même environnement bruité, mais cette fois le bruit additif n'est plus négligeable. Cette expérience a pour but de vérifier les performances lorsqu'il s'agit d'une interaction entre un seul locuteur et le robot dans un environnement bruité. Pour cette expérience, le signal séparé est ramené dans le domaine temporel et passé au travers une fenêtre de lissage pour en extraire finalement le spectre. Cette procédure redondante est utilisée simplement pour permettre de valider les résultats en écoutant chaque segment identifié.

La position des locuteurs est illustrée à la figure 5.13. La figure 5.14 illustre les résultats obtenus. Les performances avec les indices pondérés avec masques se situent entre 74.2% et 100.0%, pour une moyenne de 90.6%. Un taux semblable est obtenu avec le système présenté par Kim et al. [24] (environ 90% de bonnes identifications non-pondérées avec une base de données de 30 locuteurs), mais ces résultats sont valides uniquement pour un environnement silencieux. Les expériences actuelles sont effectuées en environnement bruité, ce qui correspond davantage à un scénario réaliste.

Il est intéressant de constater que les performances ne sont pas les mêmes d'un locuteur à l'autre. Ceci s'explique par le fait que certains locuteurs possèdent des caractéristiques vocales semblables. On constate ainsi que le locuteur D partage plusieurs de ses caractéristiques avec les locuteurs A, B et C. Par contre, les locuteurs A, B et C possèdent pour leurs parts des caractéristiques uniques qui ne sont pas présentes chez le locuteur D. Ceci fait en sorte que le locuteur D peut être plus facilement reconnu comme étant le locuteur A, B ou C tandis que l'inverse est moins probable.

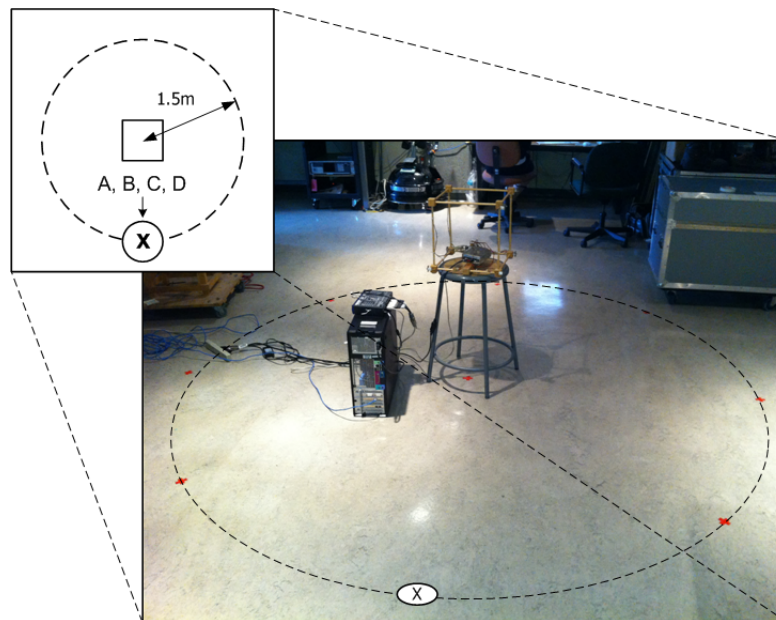


Figure 5.13 Positions des locuteurs pour le scénario statique

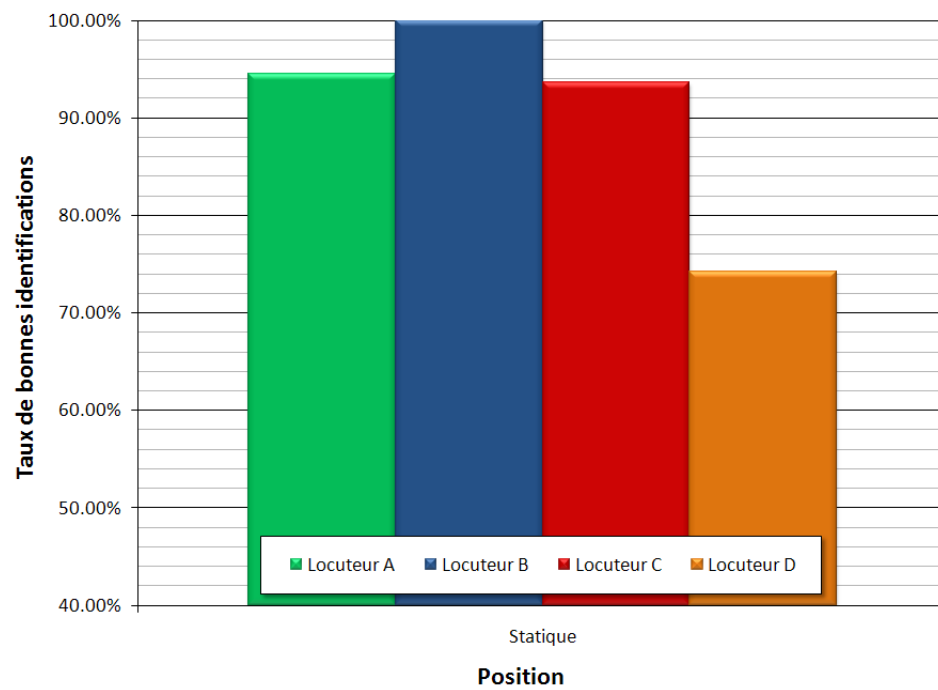


Figure 5.14 Performances avec les indices pondérés pour chaque locuteur avec masques dans un contexte d'interaction statique

5.3 Interaction dynamique

Finalement, une expérience d'interaction dynamique est introduite et consiste à utiliser le système dans une situation de conversation naturelle entre les locuteurs mâles présents qui ont participé à l'expérimentation d'interaction statique. Les mêmes modèles de locuteurs utilisés pour l'interaction statique sont employés pour cette expérience. Durant cette conversation, chaque locuteur est positionné à une distance de 1.5 mètres du cube et un espace de 90 degrés est présent entre chacun. Cette disposition est illustrée à la figure 5.15. Il est à noter que les coordonnées angulaires sont définies dans le sens horaire. Ces locuteurs parlent à tour de rôle de différents sujets avec un fil de discussion naturel. Ainsi, il arrive fréquemment que plusieurs locuteurs parlent simultanément. En effet, lorsqu'un locuteur principal dialogue, les autres locuteurs vont souvent prononcer quelques mots pour effectuer une approbation. De plus, le fil de la discussion comprend souvent des chevauchements entre la fin de la phrase d'un premier locuteur et le début de celle d'un deuxième locuteur. Cette réalité met à l'épreuve le système car les segments de parole d'un locuteur sont souvent contaminés par d'autres locuteurs, affectant ainsi les performances d'identification.

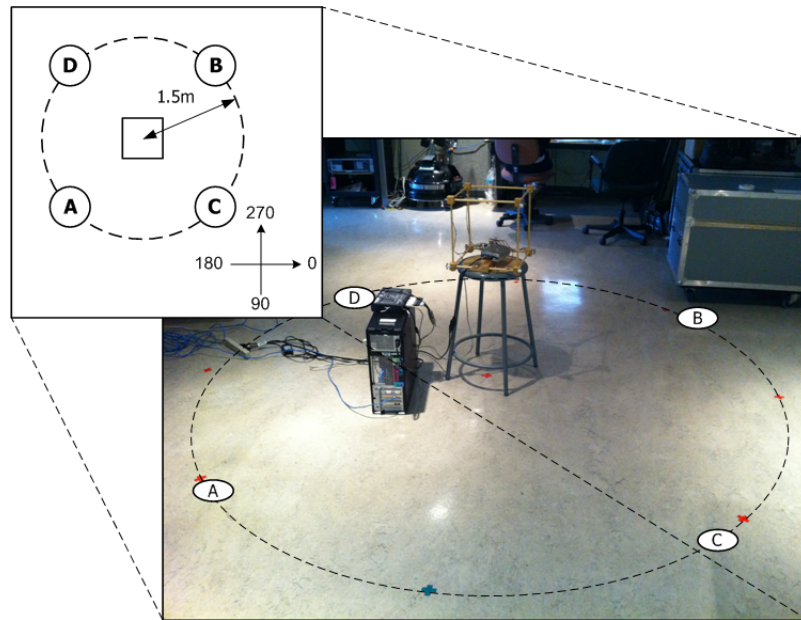


Figure 5.15 Positions des locuteurs pour le scénario dynamique

Pour cette expérience, le taux d'adaptation μ de l'équation 3.77 pour la séparation est fixé à zéro. En effet, lorsqu'une seconde source apparaît et est séparée, l'effet adaptatif de la séparation qui vise à minimiser la corrélation croisée entre les deux sources modifie considérablement le spectre de la source initialement séparée. Ce changement abrupt modifie

les caractéristiques du canal et affecte l'estimation des bruits convolutif et additif. Les modèles modifiés dynamiquement ne représentent alors pas fidèlement les conditions de l'environnement, ce qui entraîne une détérioration importante des performances. Il a été observé que les performances avec les indices pondérés et les masques se situaient alors sous le seuil de 50%. En fixant μ à zéro, la séparation du signal de plusieurs locuteurs est moins prononcée, ce qui diminue les performances de reconnaissance de la parole à plusieurs locuteurs qui parlent simultanément mais améliorent les performances d'identification des locuteurs.

Les taux de succès au niveau de l'identification en mode d'interaction dynamique sont résumés à la figure 5.16 et au tableau 5.4 en fonction des trois longueurs de segments étudiés. La longueur des segments est déterminée selon la durée des caractéristiques vocales valides (sans les périodes de silence). Les résultats détaillés sont présentés à l'annexe A.

Tableau 5.4 Durée des segments et taux de bonnes identifications

Longueur	Durée (secs)	Taux de bonnes identifications		
		Minimum	Maximum	Moyenne
Courte	1.0	42.6%	71.8%	58.3%
Moyenne	2.0	45.9%	100.0%	72.8%
Longue	3.0	49.9%	100.0%	81.4%

Ce bilan démontre clairement qu'une plus longue durée des segments à identifier entraîne de meilleures performances, tel qu'anticipé. Il est intéressant de noter que les performances ne sont pas similaires à celles obtenues pour le scénario statique en ce qui concerne le taux de bonnes identifications selon le locuteur. En effet, le locuteur D possède le taux d'identifications le plus élevé des locuteurs dans le scénario dynamique et le moins élevé dans le scénario statique. Les locuteurs effectuent la lecture d'un texte durant l'entraînement et le scénario statique tandis qu'ils parlent naturellement durant le scénario dynamique. Les résultats semblent démontrer que la prononciation des phonèmes n'est pas la même pour ces deux scénarios et expliqueraient cette disparité dans les performances entre ces expériences. En plus d'être reliées à ce phénomène, les faibles performances pour le locuteur A semblent être causées par des segments de parole plus courts chez ce locuteur.

Il est difficile de comparer les résultats obtenus pour ce scénario à d'autres résultats existants car aucune expérience similaire n'a été recensée dans la littérature. Cette expérience suggère qu'il est possible d'identifier pour la majorité des cas le locuteur malgré un environnement bruyant et le chevauchement entre les locuteurs. Cependant, pour une identification robuste, la durée des caractéristiques vocales valides devraient être d'au moins trois sec-

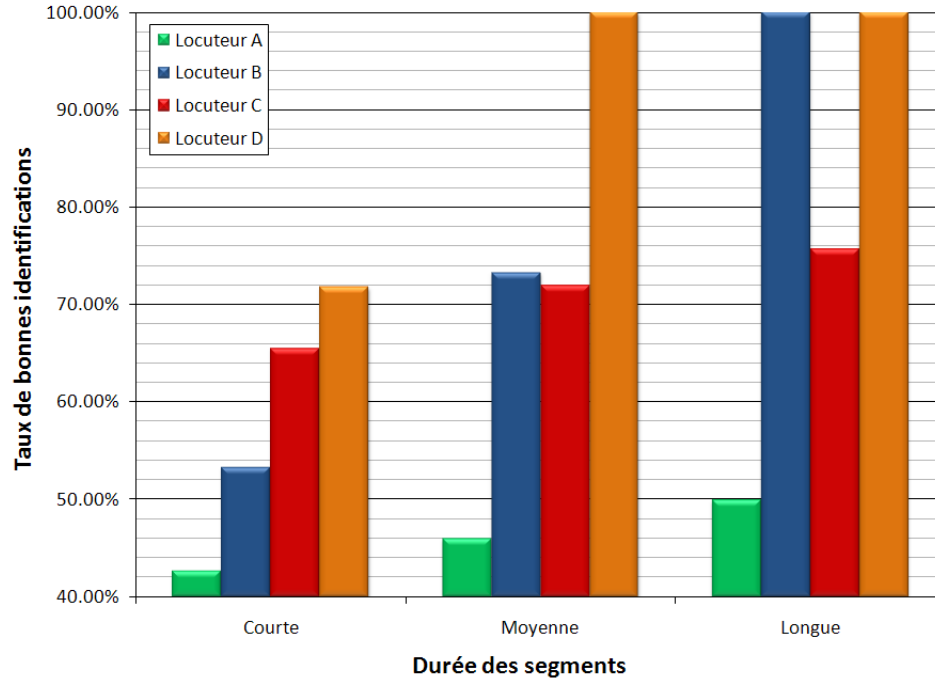


Figure 5.16 Performances avec les indices pondérés pour chaque locuteur avec masques dans un contexte d'interaction dynamique

ondes. Cette durée pourrait également être augmentée, mais dans ce cas de nombreux segments ne seraient pas analysés étant donnée la courte durée des phrases dans une discussion naturelle. En effet, il a été mesuré à partir d'une base de données qu'une phrase dans une discussion naturelle contenait en moyenne 7.7 mots [2].

5.4 Considérations temps réel

Jusqu'à présent, les performances ont été présentées en terme de taux de bonnes identifications de locuteurs. Cependant, le temps d'exécution de l'algorithme doit également être pris en compte car ce système doit pouvoir fonctionner en temps réel sur un robot.

Le système ManyEars est implémenté en langage C et fonctionne sur un GPP. Des instructions extension pour flux SIMD (*Streaming SIMD Extension* (SSE)) sont utilisées pour accélérer les calculs sur des vecteurs de données. ManyEars est exécuté sur un seul fil d'exécution séquentiel. Dans le cas présent, un processeur Intel i7 cadencé à 2.93 GHz est utilisé pour effectuer les calculs. Pour les scénarios actuels, c'est-à-dire lorsqu'un locuteur principal est actif sur une longue durée et que les autres locuteurs sont actifs de manière sporadique, le système ManyEars utilise environ 80% des ressources d'un seul cœur du GPP.

Pour les expériences précédentes, le système WISS est exécuté avec Matlab qui est un langage interprété. L'environnement Matlab est choisi pour ses outils de développement bien que l'exécution soit moins rapide qu'un logiciel compilé en langage C. Malgré cela, cet algorithme utilise tout de même seulement 10% des ressources d'un seul cœur du GPP. L'ajout d'un système de reconnaissance de locuteurs sur un robot mobile est donc attrayant car la charge supplémentaire de calculs est faible comparativement aux calculs nécessaires pour le système ManyEars. La prochaine phase de développement prévoit l'intégration du système WISS en langage C, ce qui devrait réduire la charge de calcul davantage.

La reconnaissance de locuteurs peut également s'effectuer parallèlement à l'exécution du système ManyEars. En effet, le système ManyEars se doit d'être synchrone avec l'échantillonnage des microphones pour éviter un trop grand délai et un débordement potentiel des tampons. Les trames en provenance des microphones sont généralement mises temporairement en mémoire dans un tampon circulaire avant d'être traitées par le système ManyEars. Une fois que le spectre des signaux séparés est multiplié par le banc de filtre, ces valeurs peuvent être mises en mémoire dans un autre tampon. Lorsqu'une quantité suffisante de trames est obtenue, l'algorithme de reconnaissance peut être démarré sur un fil d'exécution différent de celui utilisé par le système ManyEars. Les calculs nécessaires pour la reconnaissance n'affectent donc pas les performances en temps réel du système ManyEars. Un aperçu de la structure proposée est illustré à la figure 5.17.

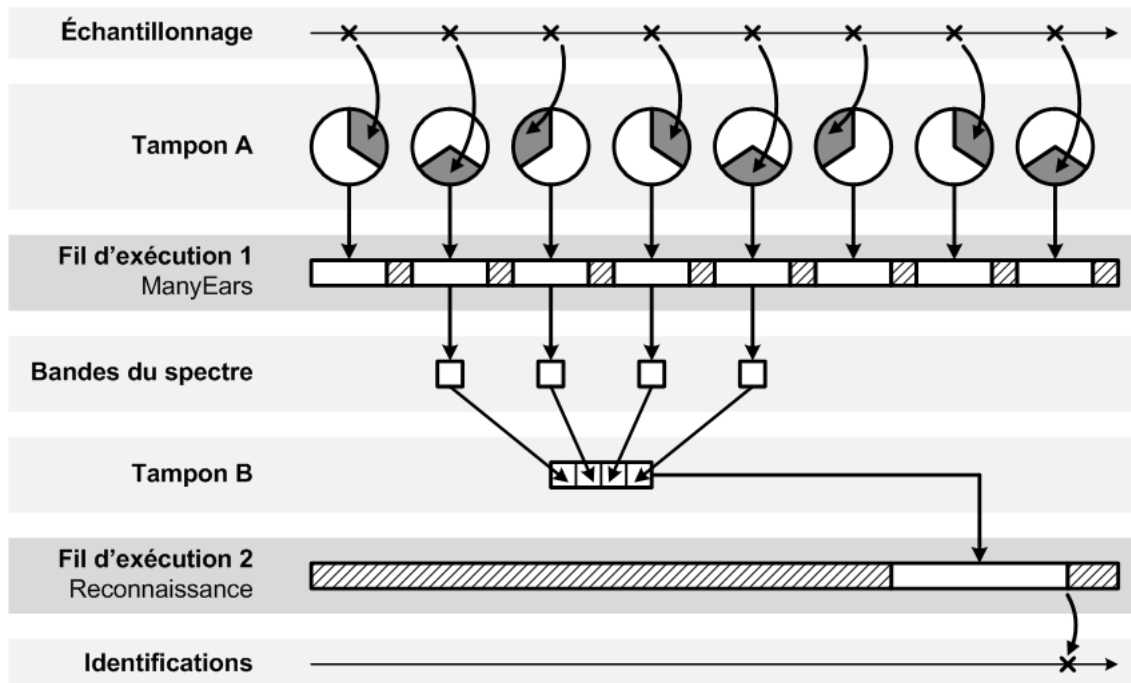


Figure 5.17 Fils d'exécution pour un scénario en temps réel

5.5 Discussion

Le système WISS implémenté démontre qu'il est possible d'effectuer une reconnaissance de locuteurs dans un milieu bruité. Lorsque la position angulaire change, les performances restent essentiellement les mêmes. Ceci indique clairement que le système est en mesure de reconnaître les locuteurs peu importe leur orientation par rapport au robot. Cette caractéristique est importante car elle permettra au robot d'interagir sans avoir à continuellement s'orienter pour faire face à la personne qui lui parle si cette dernière se déplace de temps à autre. Par contre, le taux de reconnaissance diminue lorsque le locuteur s'éloigne du robot, ce qui est normal puisque le SNR diminue dans ce cas. Le robot pourrait alors tenter de garder une distance appropriée avec le locuteur.

L'utilisation de masques démontre clairement des performances supérieures par rapport à un système sans masques, en particulier dans des cas où le SNR est faible. Cette caractéristique est particulièrement intéressante car les masques sont également utilisés dans un contexte de reconnaissance de la parole, ce qui permet deux applications différentes avec un prétraitement commun. L'utilisation d'un indice de confiance permettra également d'améliorer les performances. Dans le cas actuel, cette caractéristique peut sembler moins essentielle mais prend tout son sens dans un contexte de fusion pour lequel une reconnaissance des individus combinant plusieurs modalités est requise.

Malgré les bonnes performances obtenues, il faudrait trouver des manières de les améliorer lorsque le SNR diminue. Ce constat représente un défi de taille car le robot doit être en mesure d'interagir dans un milieu bruité. Plusieurs solutions sont possibles pour remédier à ce problème. Tout d'abord, il est démontré que le signal obtenu à l'aide d'un formateur de faisceaux possède un SNR similaire à celui du signal obtenu à l'aide de la somme de l'énergie de chaque microphone. Ceci démontre clairement que le formateur de faisceaux actuel est optimisé pour réduire le bruit directionnel mais pas le bruit ambiant additif. L'utilisation d'un formateur de faisceaux qui vise à minimiser le bruit ambiant, tel que proposé en [55], pourrait augmenter le SNR. L'utilisation d'un système avec un plus grand nombre de microphones permettrait également d'améliorer le SNR. Un tel système pourrait permettre d'entraîner des modèles de locuteurs à une distance de plusieurs mètres du robot. Ceci permettrait au robot d'entraîner facilement et rapidement des modèles lorsqu'il rencontre de nouveaux individus dans son environnement. Ceci pourrait également permettre de reconnaître plusieurs locuteurs qui parlent simultanément. Cependant, la charge de calcul associée à un système avec autant de microphones augmenterait significativement. De plus, l'augmentation du nombre de capteurs représente en soi un défi important au niveau du transfert de l'information vers la mémoire du GPP, étant donnée la large bande

passante requise. L'utilisation de codecs pour compresser ces signaux semble alors être une option intéressante. Une architecture dédiée pour un système doté d'un grand nombre de microphones a d'ailleurs été proposée [50].

Dans le système proposé, les caractéristiques vocales utilisées extraient l'information statique de chaque bande de filtres. L'utilisation de paramètres dynamiques (la première et la seconde dérivée des paramètres statiques) pourraient potentiellement améliorer les performances. Cette hypothèse devra être vérifiée dans le cadre de travaux futurs. De plus, la modélisation des caractéristiques vocales par un GMM plutôt qu'une quantification vectorielle pourrait également bonifier le taux de reconnaissance. Cette piste devra être investiguée dans le cadre de travaux futurs.

Il a été démontré que de meilleures performances sont obtenues dans un environnement bruité lorsqu'on travaille dans le domaine de la reconnaissance de la parole puisqu'il est possible de réaliser une reconnaissance satisfaisante de la parole sur plusieurs sources simultanées [46]. Toutefois, la reconnaissance de locuteurs n'est possible que pour un seul locuteur à la fois avec le système WISS. Présentement, la reconnaissance de locuteurs s'effectue en se basant sur les caractéristiques vocales acoustiques, c'est-à-dire la représentation spectrale de chaque phonème. C'est pour cette raison que tout changement spectral causé une interférence provenant d'un autre locuteur entraîne une détérioration rapide des performances. Dans le cas de la reconnaissance de la parole, chaque phonème est modélisée à partir d'un vaste ensemble d'enregistrements de plusieurs locuteurs effectués dans différents milieux bruités et non bruités. Ceci permet d'améliorer la robustesse des algorithmes de reconnaissance de phonèmes en situation de bruits ambiants. Par conséquent, une reconnaissance de locuteurs qui utilise une information de plus haut niveau, c'est-à-dire la combinaison des phonèmes propre à chaque locuteur, permet de meilleures performances en milieu bruité [56]. Le principal inconvénient de cette technique est que l'entraînement des modèles de locuteurs nécessite préalablement un modèle pour chaque phonème. Ces modèles sont générés à partir d'imposantes bases de données de parole pour plusieurs langues différentes. Malgré la complexité de cette nouvelle approche, elle semble devenir de plus en plus un incontournable puisque l'amélioration du SNR demeure limitée étant donnée la présence constante de bruits ambiants dans un environnement dynamique où est appelé à évoluer un robot mobile.

CHAPITRE 6

CONCLUSION

Ce mémoire présente le système de reconnaissance de locuteurs WISS qui se combine au système ManyEars conçu pour un robot mobile. Le système WISS utilise des caractéristiques MFCC dans le domaine spectral pour extraire les informations propres à chaque locuteur. Durant l'entraînement, un modèle est généré grâce à la technique de quantification vectorielle avec des segments de parole enregistrés dans un milieu non bruité. Par la suite, durant l'étape d'identification, des caractéristiques vocales et des masques sont générés. Grâce à une estimation des bruits additifs et convolutifs, les modèles des locuteurs sont mis à jour à l'aide de la technique PMC. Ceci permet d'adapter les modèles entraînés dans des conditions idéales à l'environnement bruité dans lequel le robot interagit avec le locuteur. Les caractéristiques obtenues sont ensuite comparées avec les modèles mis à jour en utilisant les masques pour pondérer les bandes en fonction du niveau de bruit mesuré. Cette étape permet de retourner un pointage qui détermine le niveau de similitude entre les caractéristiques du locuteur à identifier et les modèles précédemment entraînés. Un indice de confiance est également généré pour pondérer la validité de chaque identification de locuteur.

Les résultats présentés au chapitre 5 démontrent que le système atteint un taux pondéré de bonnes identifications se situant entre 84.3% et 95.6% pour l'expérience avec des haut-parleurs situés à 1.5 mètres des microphones, entre 74.2% et 100.0% pour l'expérience avec l'interaction statique et entre 42.6% et 100.0% pour l'expérience avec l'interaction dynamique (pour des moyennes de 58.3%, 72.8% et 81.4% pour des segments d'une durée de une, deux et trois secondes respectivement), et que l'utilisation des masques joue un rôle important dans l'atteinte de ces performances. Bien que ces performances se détériorent légèrement lors de l'expérience en interaction dynamique avec faible SNR et superpositions de segments de parole entre plusieurs locuteurs, le système démontre une robustesse intéressante face au bruit ambiant dans un contexte d'un environnement dynamique. De plus, ce système de reconnaissance peut être implanté en temps réel sur un GPP moderne à l'aide de deux fils d'exécution, ce qui est particulièrement approprié avec les processeurs multi-cœurs.

La réalisation de ce système a permis d'identifier l'importance d'améliorer le SNR pour un tel robot mobile équipé de plusieurs microphones. L'amélioration des algorithmes ainsi

que l'augmentation du nombre de microphones font partie des solutions proposées. L'utilisation de caractéristiques vocales en lien avec l'organisation des phonèmes constitue également une avenue intéressante pour l'interaction en milieu bruyé. L'introduction de caractéristiques vocales dynamiques et d'un GMM est également une piste à explorer.

Dans un futur rapproché, le système de reconnaissance de locuteur sera installé sur le robot Johnny-0 développé à l'Université de Sherbrooke [33]. Une estimation des positions éventuelles des microphones sur ce robot est illustrée à la figure 6.1. Les performances du système devront être évaluées en présence du bruit interne du robot généralement produit par ses actionneurs mécaniques. Le système sera également évalué lorsque le robot se déplace dans la pièce et que les caractéristiques du bruit et de la réverbération changent continuellement selon sa position.

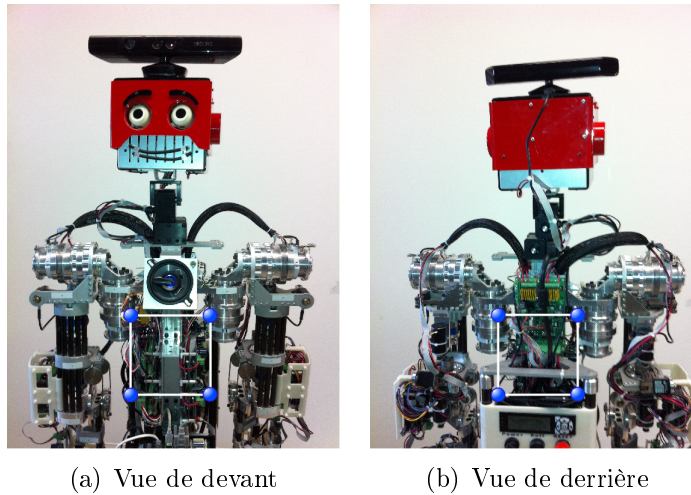


Figure 6.1 Positions des microphones sur le robot Johnny-0

ANNEXE A

INTERACTION DYNAMIQUE

Les résultats de l'expérience avec le scénario dynamique tel qu'illustré à la figure A.1 sont présentés dans les figures des sections suivantes. Les couleurs verte, bleue, rouge et orange représentent les locuteurs A, B, C et D respectivement. La position de la ou des sources est affichée selon l'azimut en degrés par rapport au temps en seconde. L'azimut est défini dans le sens horaire. Dans chaque figure, le graphique du haut représente l'identité et la position des sources sonores théoriques. Ces informations ont été obtenues suite à des tests d'écoute minutieux et une transcription manuelle des segments de parole pour chaque locuteur. Le graphique du bas représente le suivi des sources sonores et l'identification effectuée par WISS. Pour chaque source identifiée, un encadré est affiché et contient le taux de confiance pour ce segment. Lorsque l'encadré est pointillé, cela signifie que le segment correspondant est la suite du segment affiché dans la figure précédente. Les segments de couleur noire sont ignorés pour l'identification car leur durée en termes de leurs caractéristiques vocales valides (sans les périodes de silence) est inférieure à un seuil déterminé ($T_{minFeatures}$). Il est en effet difficile pour des segments très courts de normaliser le canal et une identification valide est alors peu probable.

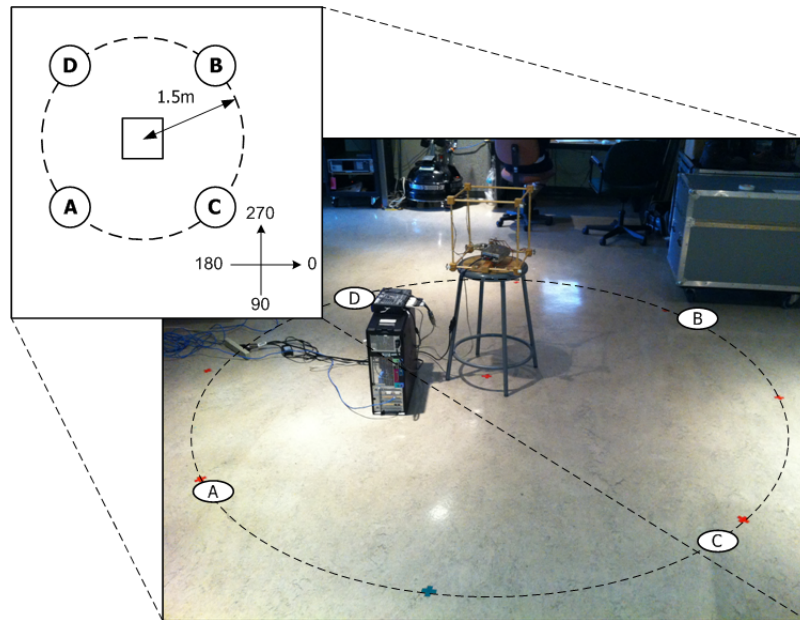


Figure A.1 Positions des locuteurs pour le scénario dynamique

A.1 Courts segments

Pour cette expérience, tous les segments qui possèdent au moins une seconde de caractéristiques valides sont identifiés. Le seuil $T_{minFeatures}$ est donc fixé à une seconde. Les résultats correspondants sont présentés dans les figures A.2 à A.13.

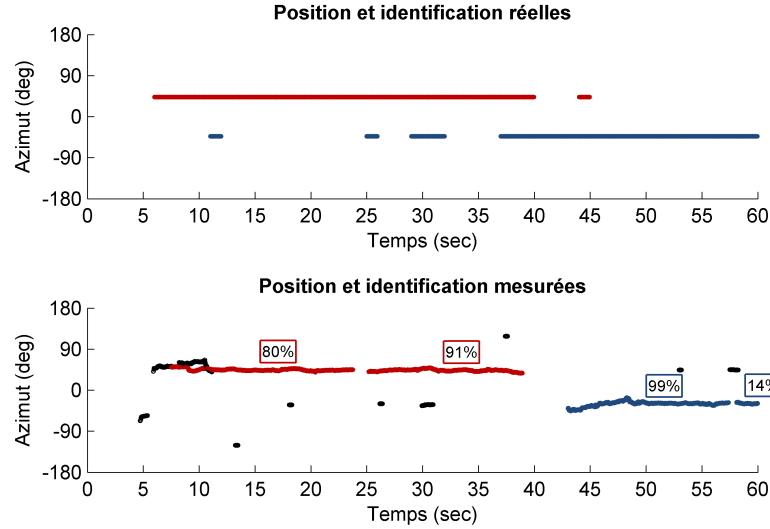


Figure A.2 Identifications pour des segments de courte durée (0-60 secs)

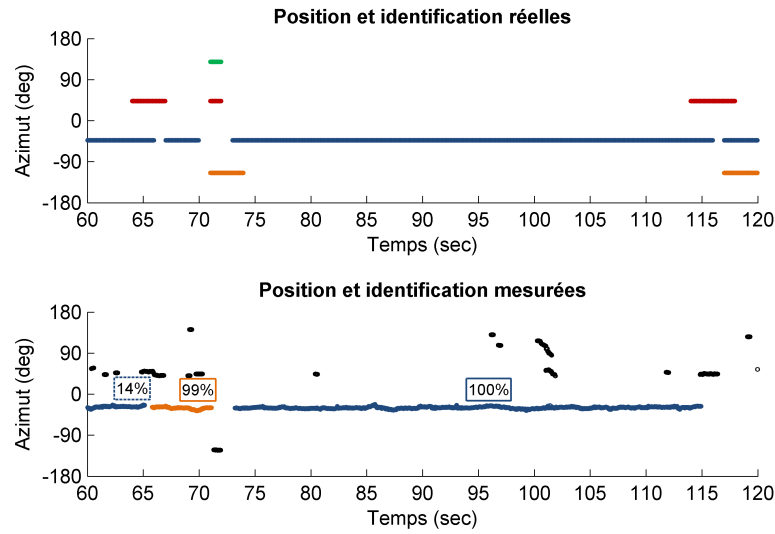


Figure A.3 Identifications pour des segments de courte durée (60-120 secs)

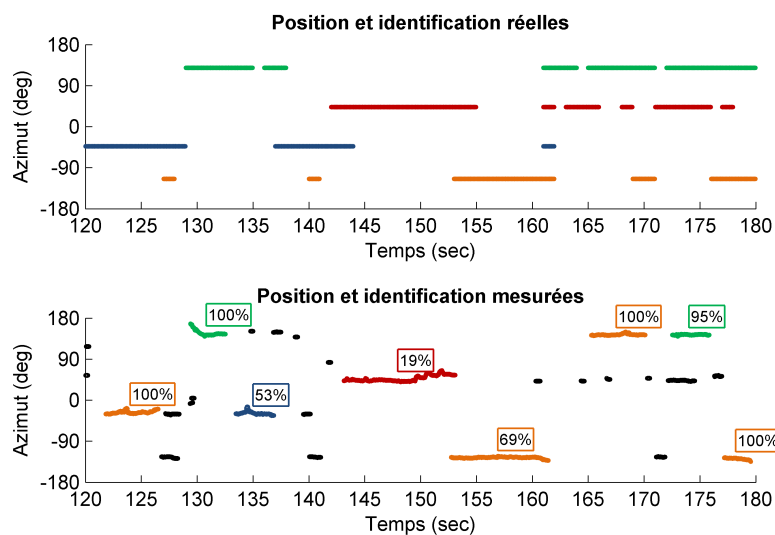


Figure A.4 Identifications pour des segments de courte durée (120-180 secs)

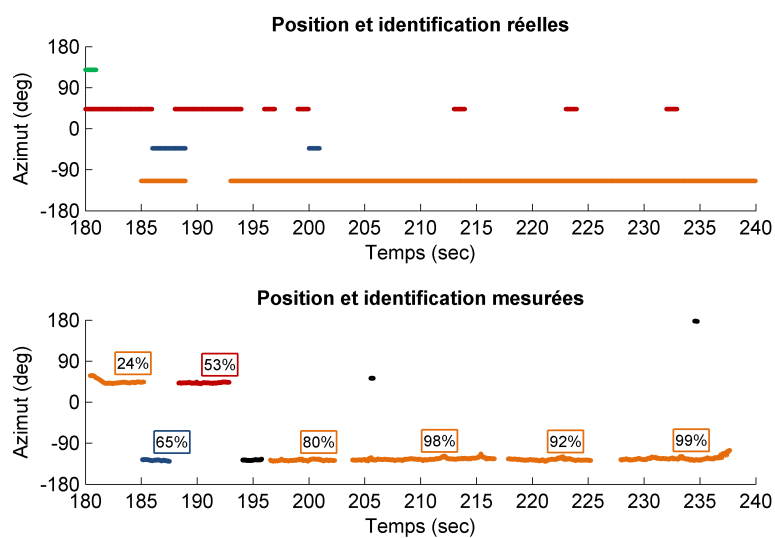


Figure A.5 Identifications pour des segments de courte durée (180-240 secs)

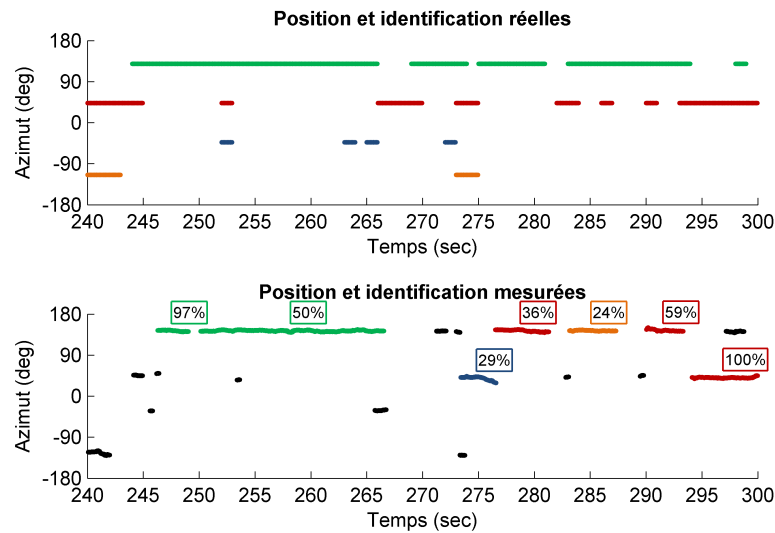


Figure A.6 Identifications pour des segments de courte durée (240-300 secs)

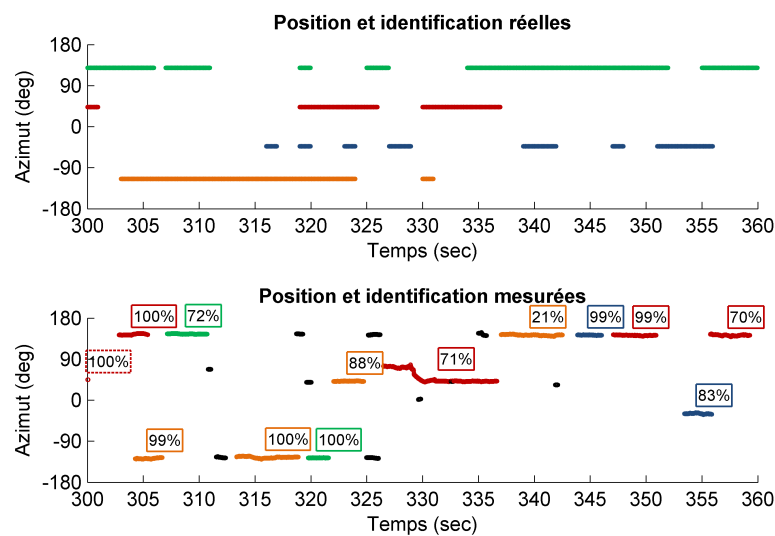


Figure A.7 Identifications pour des segments de courte durée (300-360 secs)

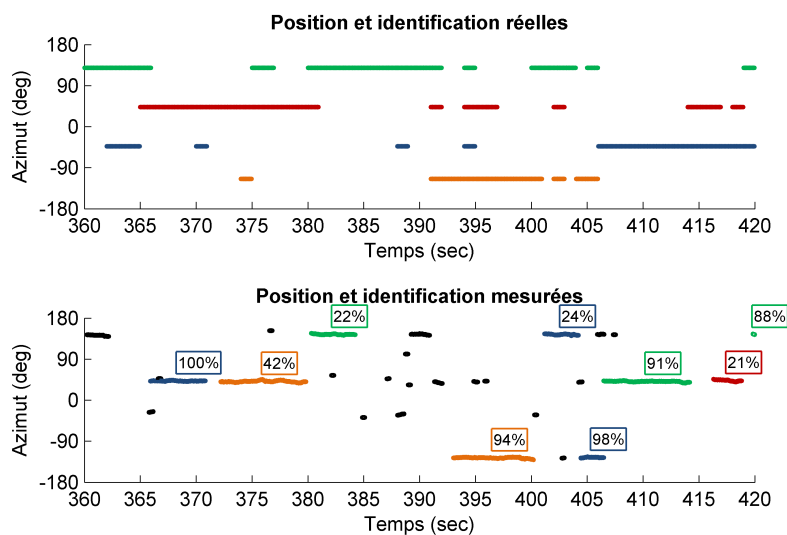


Figure A.8 Identifications pour des segments de courte durée (360-420 secs)

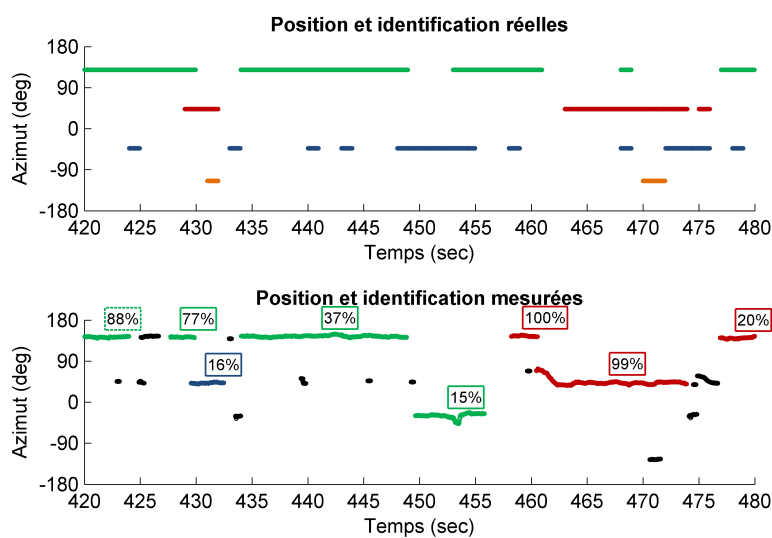


Figure A.9 Identifications pour des segments de courte durée (420-480 secs)

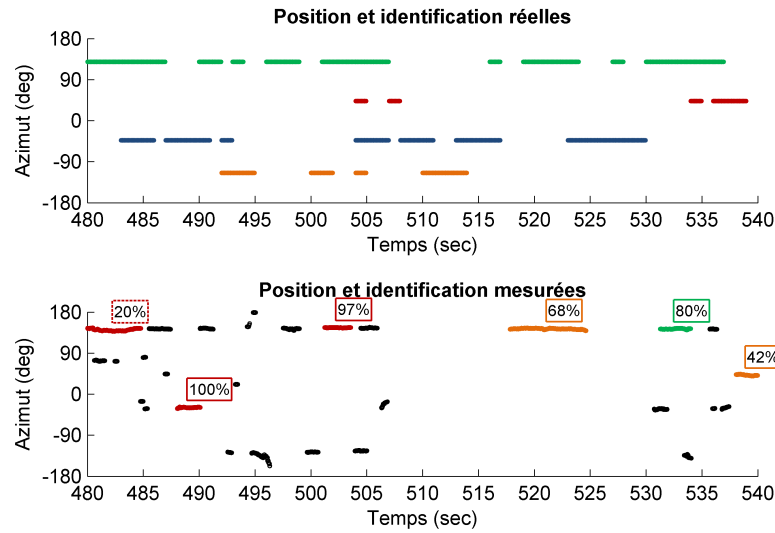


Figure A.10 Identifications pour des segments de courte durée (480-540 secs)

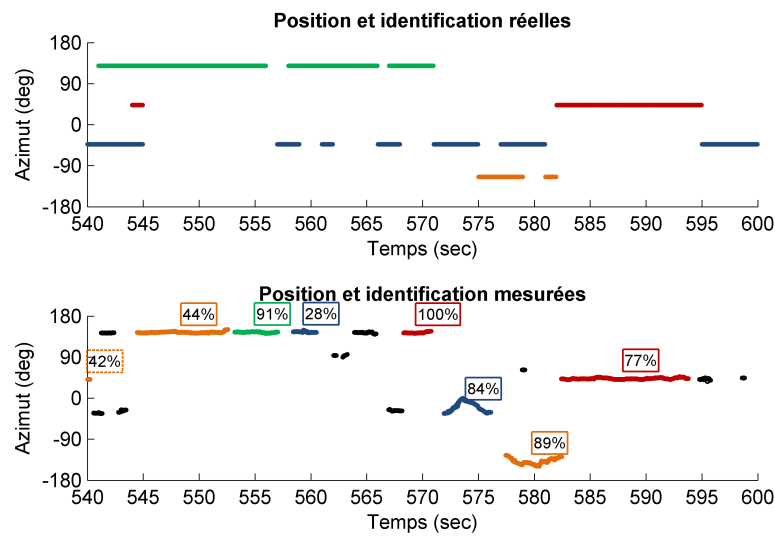


Figure A.11 Identifications pour des segments de courte durée (540-600 secs)

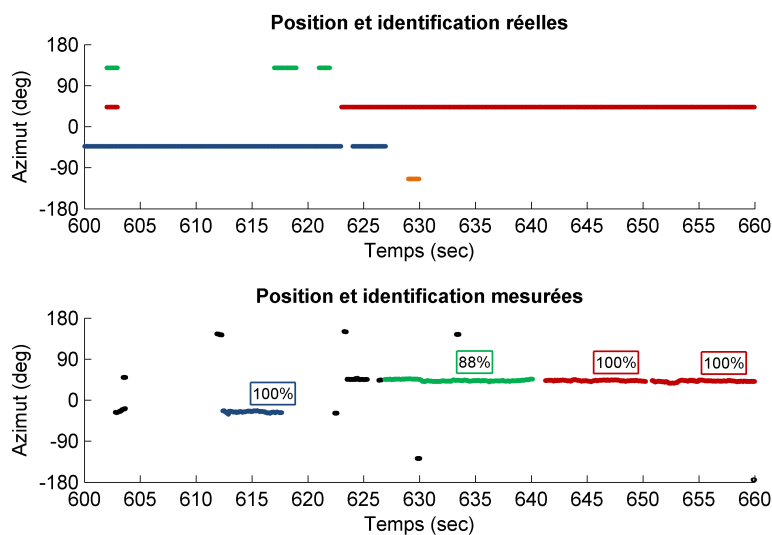


Figure A.12 Identifications pour des segments de courte durée (600-660 secs)

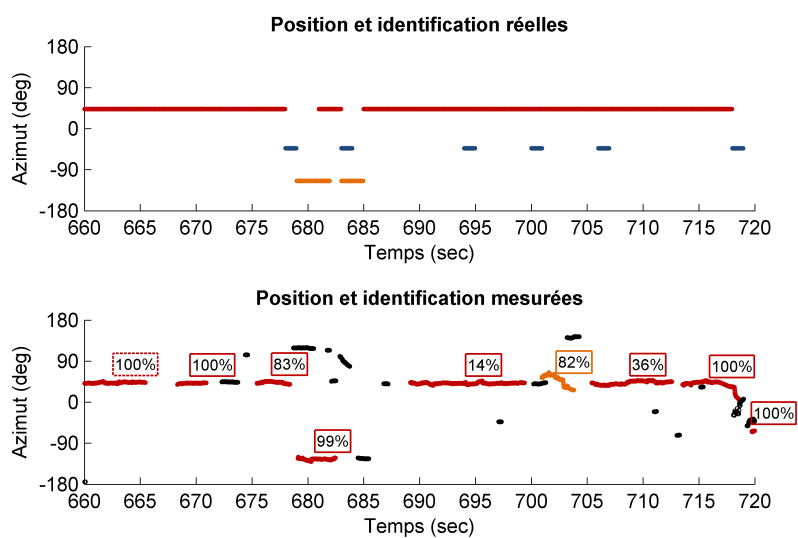


Figure A.13 Identifications pour des segments de courte durée (660-720 secs)

A.2 Moyens segments

Pour cette expérience, le seuil $T_{minFeatures}$ est fixé à deux secondes. Les résultats sont présentés dans les figures A.14 à A.25.

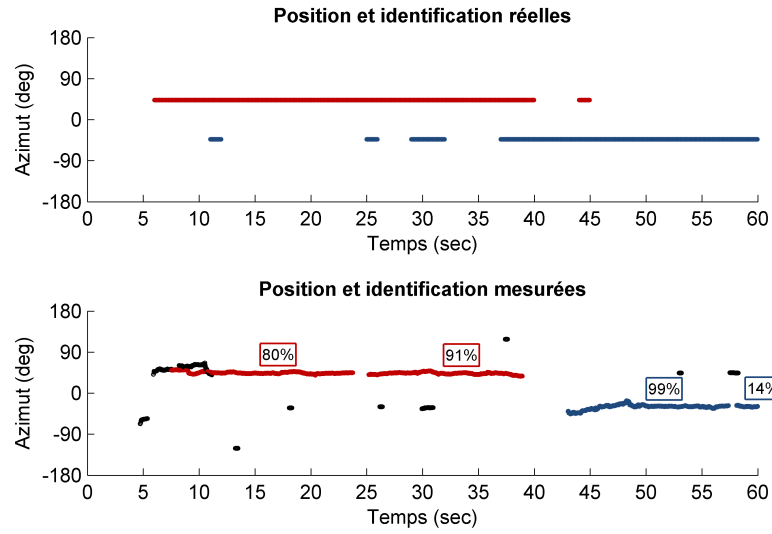


Figure A.14 Identifications pour des segments de durée moyenne (0-60 secs)

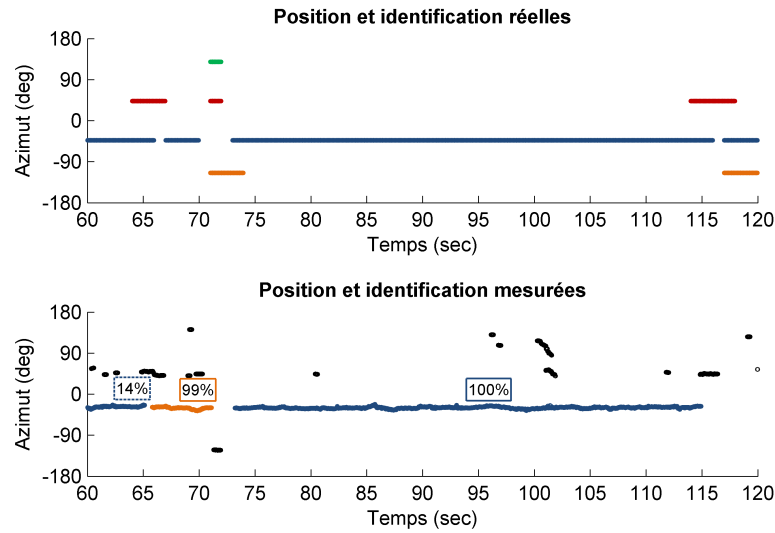


Figure A.15 Identifications pour des segments de durée moyenne (60-120 secs)

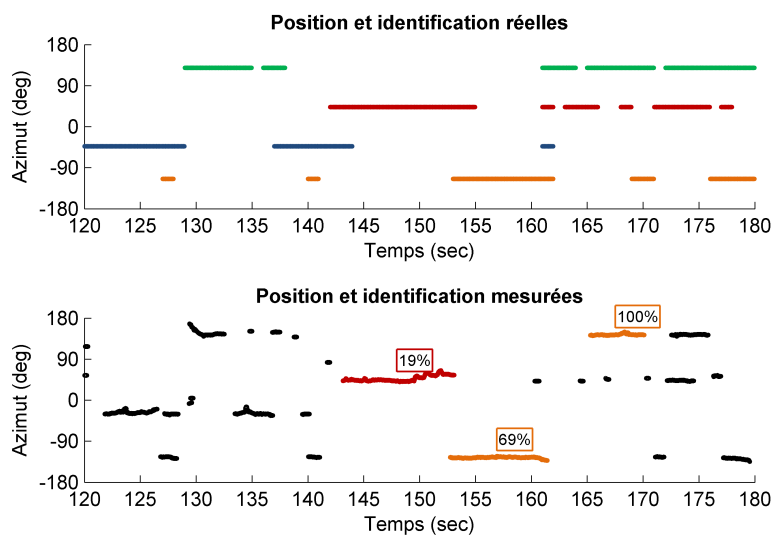


Figure A.16 Identifications pour des segments de durée moyenne (120-180 secs)

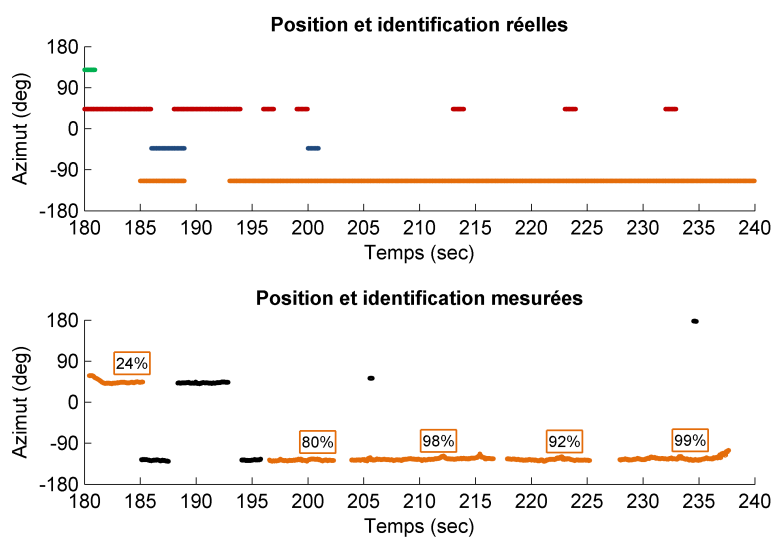


Figure A.17 Identifications pour des segments de durée moyenne (180-240 secs)

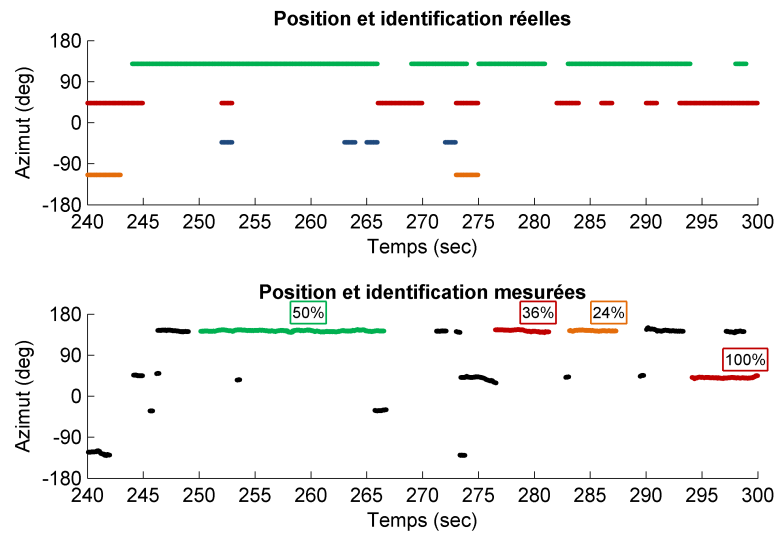


Figure A.18 Identifications pour des segments de durée moyenne (240-300 secs)

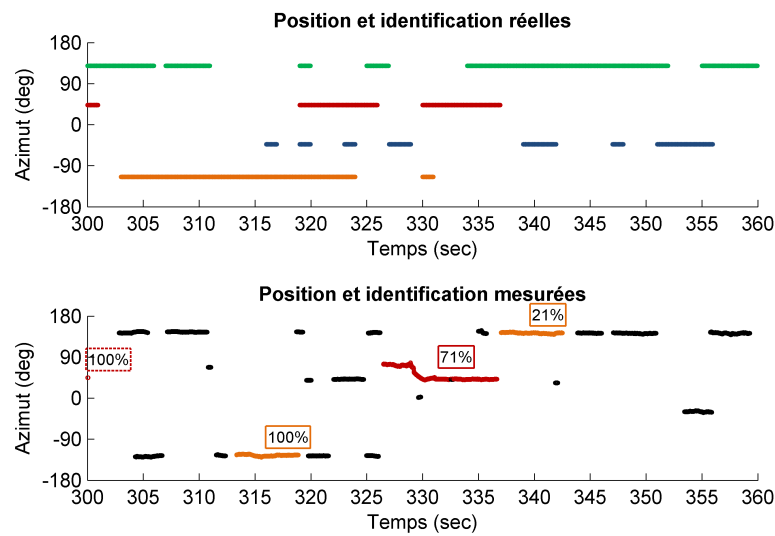


Figure A.19 Identifications pour des segments de durée moyenne (300-360 secs)

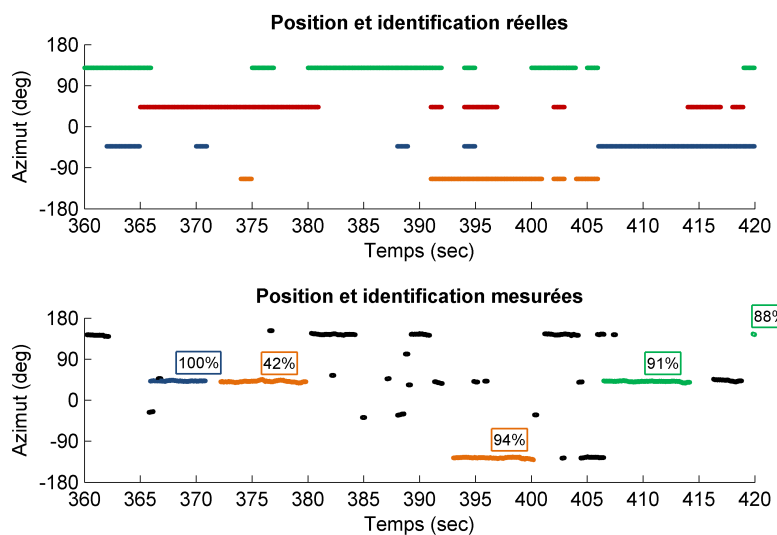


Figure A.20 Identifications pour des segments de durée moyenne (360-420 secs)

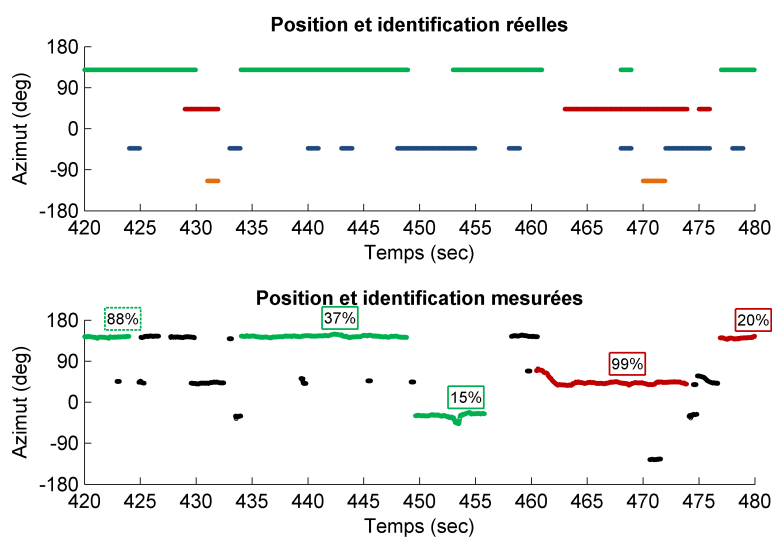


Figure A.21 Identifications pour des segments de durée moyenne (420-480 secs)

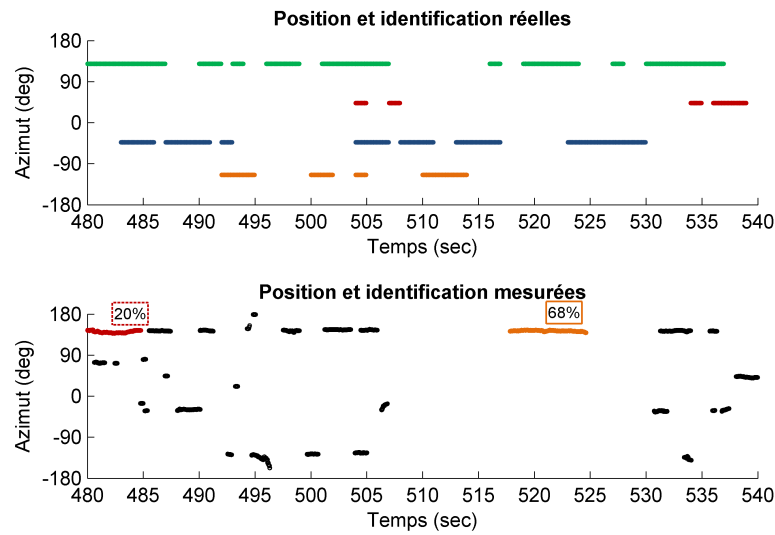


Figure A.22 Identifications pour des segments de durée moyenne (480-540 secs)

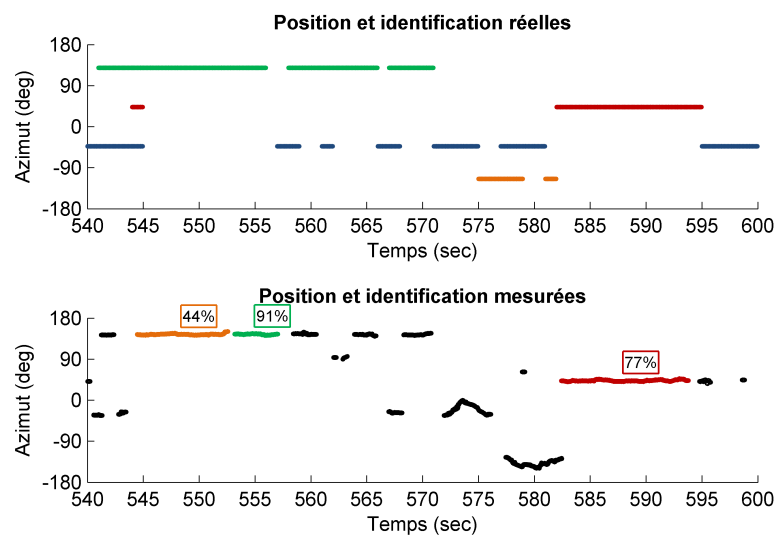


Figure A.23 Identifications pour des segments de durée moyenne (540-600 secs)

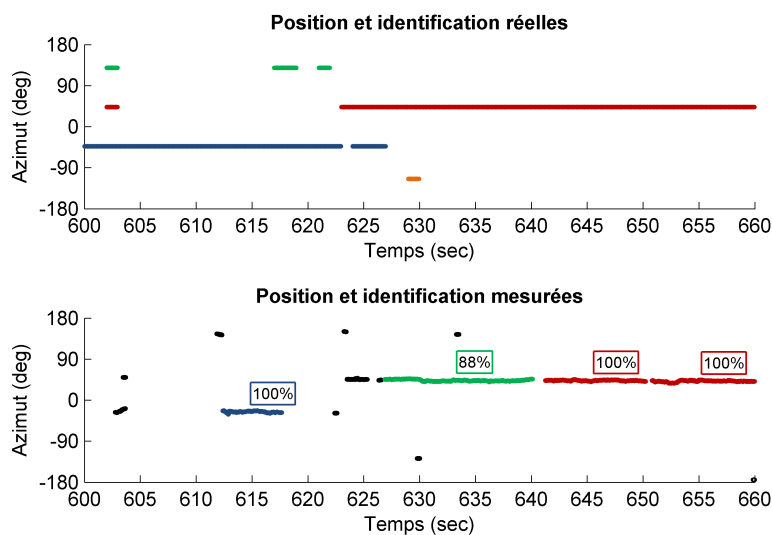


Figure A.24 Identifications pour des segments de durée moyenne (600-660 secs)

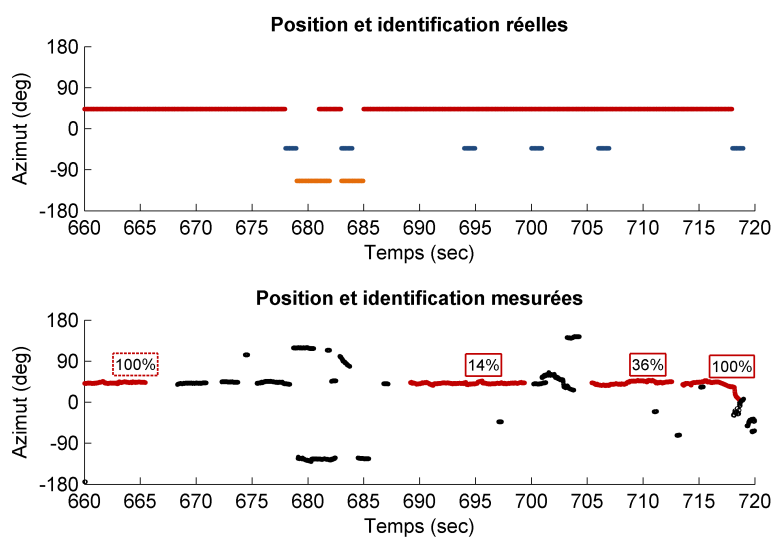


Figure A.25 Identifications pour des segments de durée moyenne (660-720 secs)

A.3 Longs segments

Pour cette expérience, le seuil $T_{minFeatures}$ est fixé à trois secondes. Les résultats sont présentés dans les figures A.26 à A.37.

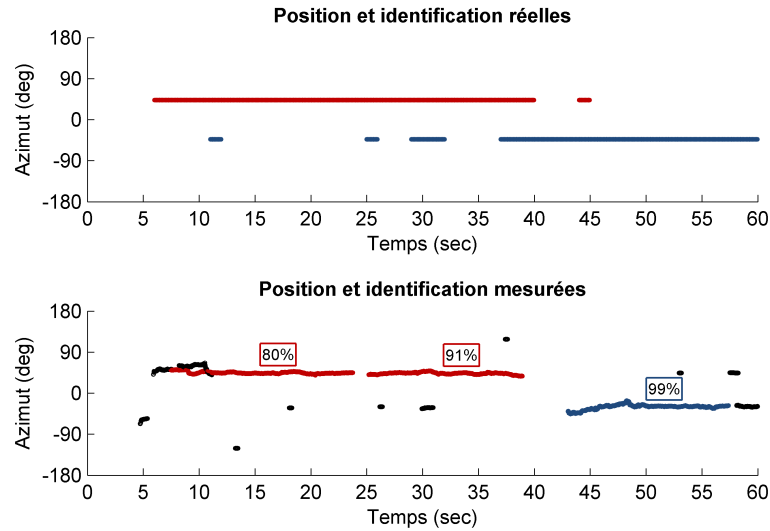


Figure A.26 Identifications pour des segments de longue durée (0-60 secs)

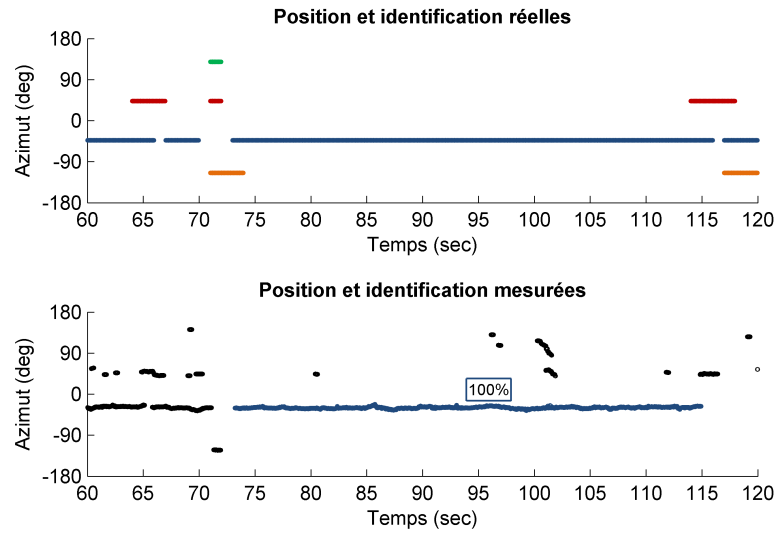


Figure A.27 Identifications pour des segments de longue durée (60-120 secs)

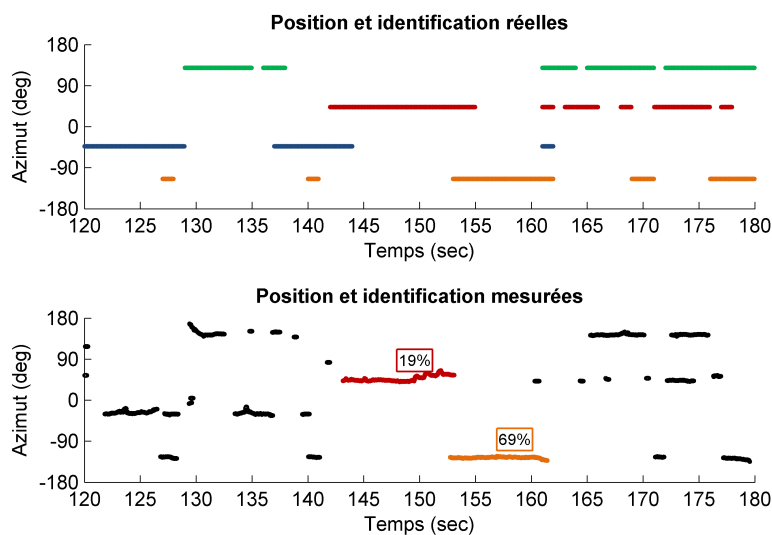


Figure A.28 Identifications pour des segments de longue durée (120-180 secs)

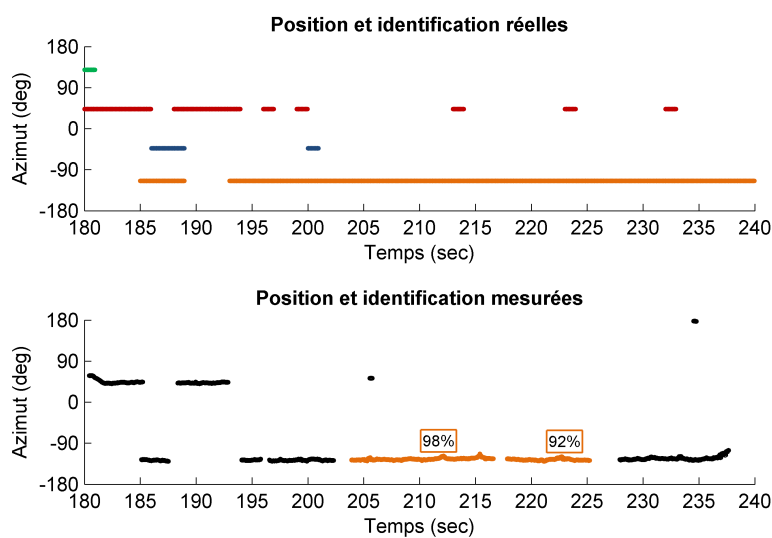


Figure A.29 Identifications pour des segments de longue durée (180-240 secs)

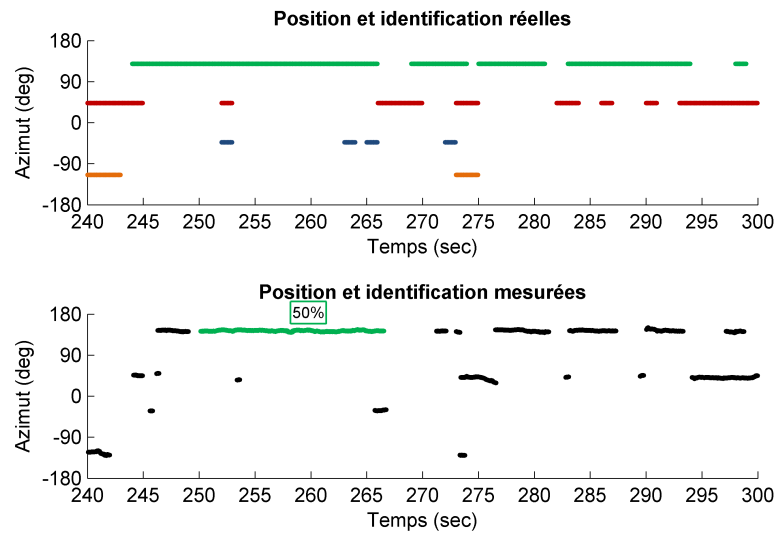


Figure A.30 Identifications pour des segments de longue durée (240-300 secs)

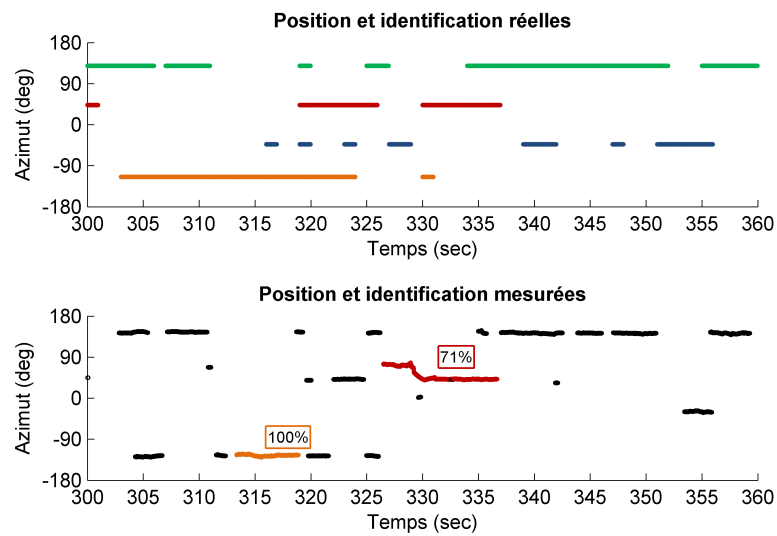


Figure A.31 Identifications pour des segments de longue durée (300-360 secs)

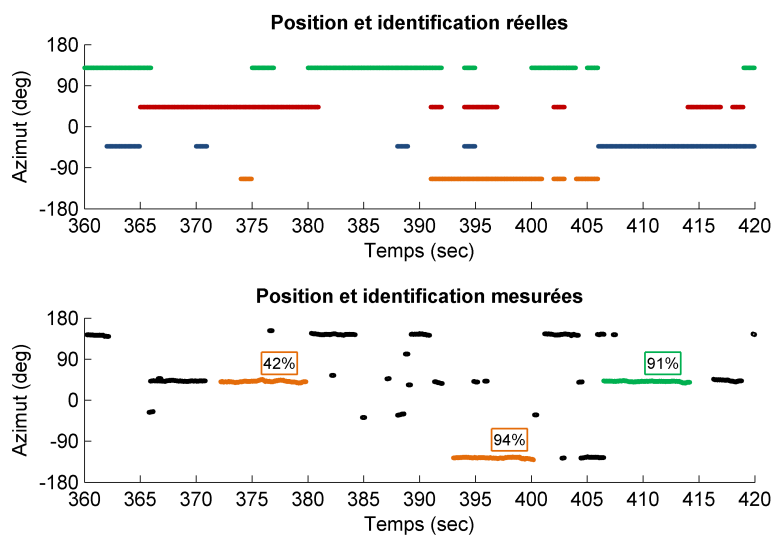


Figure A.32 Identifications pour des segments de longue durée (360-420 secs)

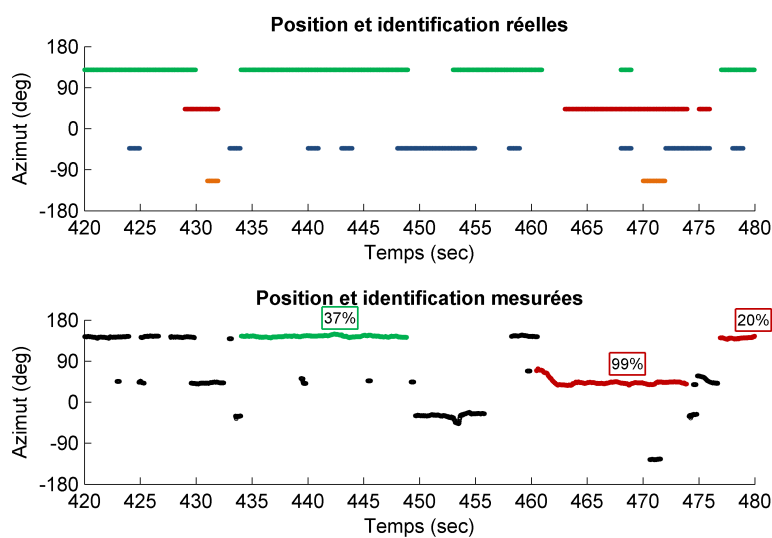


Figure A.33 Identifications pour des segments de longue durée (420-480 secs)

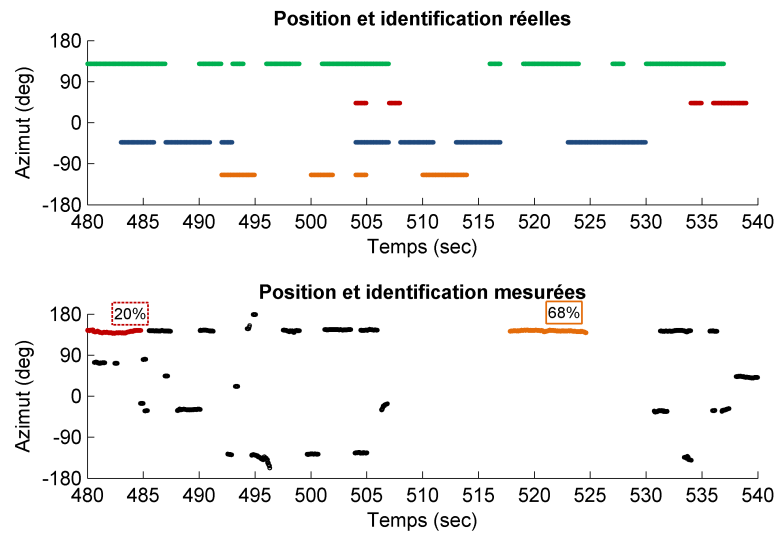


Figure A.34 Identifications pour des segments de longue durée (480-540 secs)

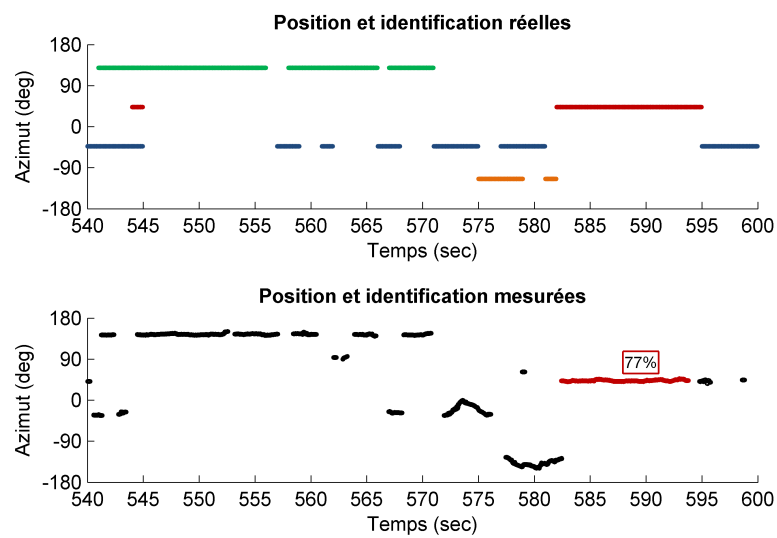


Figure A.35 Identifications pour des segments de longue durée (540-600 secs)

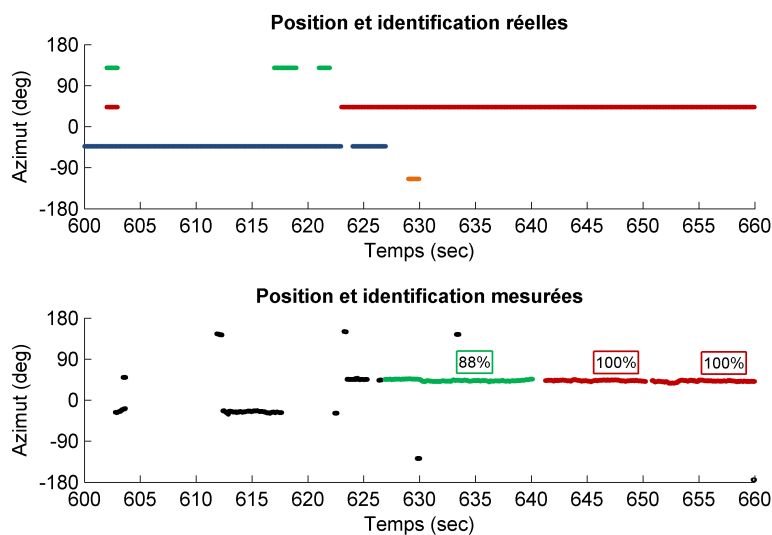


Figure A.36 Identifications pour des segments de longue durée (600-660 secs)

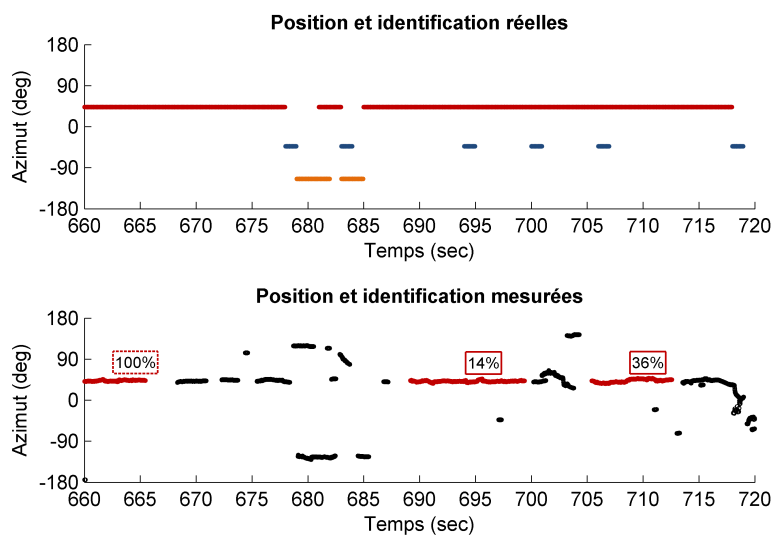


Figure A.37 Identifications pour des segments de longue durée (660-720 secs)

LISTE DES RÉFÉRENCES

- [1] Aldhaheri, R. et Al-Saadi, F. (2003) Text-independent speaker identification in noisy environment using singular value decomposition. Dans *Information, Communications and Signal Processing*, volume 3. p. 1624 – 1628.
- [2] Asami, K., Takezawa, T. et Kikui, G. (2005) Detection of topic and speech act type on utterance-by-utterance basis for conversational interfaces. *Systems and Computers in Japan*, volume 36, n° 12, p. 85 – 96.
- [3] Badali, A., Valin, J.-M., Michaud, F. et Aarabi, P. (2009) Evaluating real-time audio localization algorithms for artificial audition in robotics. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 2033 – 2038.
- [4] Ban, K.-D., Kwak, K.-C., Yoon, H.-S. et Chung, Y.-K. (2007) Fusion technique for user identification using camera and microphone in the intelligent service robots. *Proceedings of the International Symposium on Consumer Electronics*.
- [5] Briere, S., Valin, J.-M., Michaud, F. et Letourneau, D. (2008) Embedded auditory system for small mobile robots. *Proceedings of the IEEE International Conference on Robotics and Automation*, p. 3463 – 3468.
- [6] Buhmann, J. et Kuhnelt, H. (1993) Vector quantization with complexity costs. *IEEE Transactions on Information Theory*, volume 39, n° 4, p. 1133 – 1145.
- [7] Campbell, J.P., J. (1997) Speaker recognition : a tutorial. *Proceedings of the IEEE*, volume 85, n° 9, p. 1437 – 1462.
- [8] Cohen, I. et Berdugo, B. (2001) Speech enhancement for non-stationary noise environments. *Signal Processing*, volume 81, n° 11, p. 2403 – 2418.
- [9] Cole, C., Karam, M. et Aglan, H. (2008) Increasing additive noise removal in speech processing using spectral subtraction. Dans *Proceedings of the 5th International Conference on Information Technology : New Generations*. p. 1146 – 1147.
- [10] Deegan, P., Grupen, R., Hanson, A., Horrell, E., Ou, S., Riseman, E., Sen, S., Thibodeau, B., Williams, A. et Xie, D. (2008) Mobile manipulators for assisted living in residential settings. *Autonomous Robots*, volume 24, n° 2, p. 179 – 192.
- [11] Elkan, C. (2003) Using the triangle inequality to accelerate k-means. *Proceedings of the 20th International Conference on Machine Learning*, volume 1, p. 147 – 153.
- [12] Ephraim, Y. et Malah, D. (1985) Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 33, n° 2, p. 443 – 445.
- [13] Fujihara, T., Kagami, S., Sasaki, Y. et Mizoguchi, H. (2008) Arrangement optimization for narrow directivity and high SNR beam forming microphone array. *Proceedings of the IEEE Sensors*, p. 450 – 453.

- [14] Fujita, M., Kuroki, Y., Ishida, T. et Doi, T. (2003) A small humanoid robot SDR-4x for entertainment applications. *Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, volume 2, p. 938 – 943.
- [15] Gales, M. F. J. et Young, S. J. (1995) Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer Speech and Language*, volume 9, n° 4, p. 289 – 307.
- [16] Garcia, A. et Mammone, R. (1999) Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping. Dans *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1. p. 325 – 328.
- [17] Gehrig, T., Klee, U., McDonough, J., Ikbal, S., Wolfel, M. et Fugen, C. (2006) Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters. Dans *Proceedings of the 9th International Conference on Spoken Language Processing*, volume 5. p. 2594 – 2597.
- [18] Gong, Y. (2005) A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition. *IEEE Transactions on Speech and Audio Processing*, volume 13, n° 5, p. 975 – 983.
- [19] Gonzalez-Rodriguez, J., Ortega-Garcia, J., Martin, C. et Hernandez, L. (1996) Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays. Dans *Proceedings of the 4th International Conference on Spoken Language*, volume 3. p. 1333 – 1336.
- [20] Gover, J. et Huray, P. (2000) The engineer's role in averting the pending health care cost crisis. *Proceedings of the IEEE Engineering Management Society*, p. 687 – 691.
- [21] Hong, W. et Jin'gui, P. (2010) Modified MFCCs for robust speaker recognition. Dans *Proceedings of the IEEE International Conference on Intelligent Computing and Intelligent Systems*, volume 1. p. 276 – 279.
- [22] Hsu, C.-W. et Lee, L.-S. (2009) Higher order cepstral moment normalization for improved robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, volume 17, n° 2, p. 205 – 220.
- [23] Huand, X., Acero, A. et Hon, H.-W. (2001) *Spoken Language Processing : A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, New Jersey.
- [24] Ji, M., Kim, S., Kim, H. et Yoon, H.-S. (2008) Text-independent speaker identification using soft channel selection in home robot environments. *IEEE Transactions on Consumer Electronics*, volume 54, n° 1, p. 140 – 144.
- [25] Kim, H.-S. et Choi, J.-S. (2009) Sound source localization using sparse coding and som. *Proceedings of the 14th IEEE International Conference on Emerging Technologies*, p. 7.

- [26] Kinnunen, T., Kilpelainen, T. et Franti, P. (2000) Comparison of clustering algorithms in speaker identification. *Proceedings of the IASTED International Conference*, p. 222 – 227.
- [27] Kitano, H., Okuno, H., Nakadai, K., Sabisch, T. et Matsui, T. (2000) Design and architecture of SIG the humanoid : An experimental platform for integrated perception in RoboCup humanoid challenge. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, p. 181 – 190.
- [28] Kumatani, K., Gehrig, T., Mayer, U., Stoimenov, E., McDonough, J. et Wolfel, M. (2007) Adaptive beamforming with a minimum mutual information criterion. *IEEE Transactions on Audio, Speech and Language Processing*, volume 15, n° 8, p. 2527 – 2541.
- [29] Lee, J.-M., Choi, J.-S., Lim, Y.-S., Kim, H.-S. et Park, M. (2008) Intelligent and active system for human-robot interaction based on sound source localization. *Proceedings of the International Conference on Control, Automation and Systems*, p. 2738 – 2741.
- [30] Li, H.-Y., Zhao, Q.-H., Ren, G.-L. et Xiao, B.-J. (2009) Speech enhancement algorithm based on independent component analysis. Dans *Proceedings of the 5th International Conference on Natural Computation*, volume 2. Coll. of Inf. Eng., Taiyuan Univ. of Technol., Taiyuan, China, IEEE, p. 598 – 602.
- [31] Martinson, E. et Brock, D. (2007) Improving human-robot interaction through adaptation to the auditory scene. *Proceedings of the ACM/IEEE Conference on Human-Robot Interaction*, p. 113 – 120.
- [32] Michaud, F., Cote, C., Letourneau, D., Brosseau, Y., Valin, J.-M., Beaudry, E., Raievsky, C., Ponchon, A., Moisan, P., Lepage, P., Morin, Y., Gagnon, F., Giguere, P., Roux, M.-A., Caron, S., Frenette, P. et Kabanza, F. (2007) Spartacus attending the 2005 AAAI conference. *Autonomous Robots*, volume 22, n° 4, p. 369 – 383.
- [33] Michaud, F., Ferland, F., Létoirneau, D., Legault, M.-A. et Lauria, M. (2010) Toward autonomous, compliant, omnidirectional humanoid robots for natural interaction in real-life settings. *Journal of Behavioral Robotics*, volume 1, n° 1, p. 57 – 65.
- [34] Mwema, W. et Mwangi, E. (1996) A spectral subtraction method for noise reduction in speech signals. Dans *Proceedings of the 4th IEEE Africon Conference*, volume 1. p. 382 – 385.
- [35] Nakajima, H., Nakadai, K., Hasegawa, Y. et Tsujino, H. (2008) High performance sound source separation adaptable to environmental changes for robot audition. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, p. 2165 – 2171.
- [36] Navratil, J., Jin, Q., Andrews, W. et Campbell, J. (2003) Phonetic speaker recognition using maximum-likelihood binary-decision tree models. Dans *Acoustics, Speech, and Signal Processing*, volume 4. p. 796 – 9.

- [37] Parra, L. C. et Alvino, C. V. (2002) Geometric source separation : Merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, volume 10, n° 6, p. 352 – 362.
- [38] Povey, D., Chu, S. et Varadarajan, B. (2008) Universal background model based speech recognition. Dans *Acoustics, Speech and Signal Processing*. p. 4561 – 4564.
- [39] Reynolds, D. (1995) Large population speaker identification using clean and telephone speech. *IEEE Signal Processing Letters*, volume 2, n° 3, p. 46 – 48.
- [40] Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D. et Xiang, B. (2003) The SuperSID project : Exploiting high-level information for high-accuracy speaker recognition. Dans *Acoustics, Speech, and Signal Processing*, volume 4. p. 784 – 787.
- [41] Reynolds, D. et Rose, R. (1995) Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, volume 3, n° 1, p. 72 – 83.
- [42] Seddik, H., Rahmouni, A. et Sayadi, M. (2004) Text independent speaker recognition using the mel frequency cepstral coefficients and a neural network classifier. Dans *First International Symposium on Control, Communications and Signal Processing*. p. 631 – 634.
- [43] Shao, Y. et Wang, D. (2006) Robust speaker recognition using binary time-frequency masks. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, p. 645 – 648.
- [44] Shimoda, T., Nakashima, T., Kumon, M., Kohzawa, R., Mizumoto, I. et Iwai, Z. (2006) Spectral cues for robust sound localization with pinnae. *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, p. 386 – 391.
- [45] Valin, J.-M., Michaud, F. et Rouat, J. (2007) Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, volume 55, n° 3, p. 216 – 228.
- [46] Valin, J.-M., Yamamoto, S., Rouat, J., Michaud, F., Nakadai, K. et Okuno, H. G. (2007) Robust recognition of simultaneous speech by a mobile robot. *IEEE Transactions on Robotics*, volume 23, n° 4, p. 742 – 752.
- [47] Wang, G., Wang, X. et Zhao, X. (2008) Speech enhancement based on a combined spectral subtraction with spectral estimation in various noise environment. Dans *Proceedings of the International Conference on Audio, Language and Image Processing*. p. 1424 – 1429.
- [48] Wang, L., Kitaoka, N. et Nakagawa, S. (2006) Robust distant speech recognition by combining multiple microphone-array processing with position-dependent CMN. *Eurasip Journal on Applied Signal Processing*, volume 2006, p. 1 – 11.

- [49] Ward, D. B. et Williamson, R. C. (2002) Particle filter beamforming for acoustic source localization in a reverberant environment. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, p. II/1777 – II/1780.
- [50] Weinstein, E., Steele, K., A., A. et Glass, J. (2007) Loud : A 1020-node microphone array and acoustic beamformer. *Proceedings of the International Congress on Sound and Vibration*.
- [51] Wong, L. P. et Russell, M. J. (2001) Speaker verification under additive noise conditions with non-stationary SNR using PMC. Dans *The Speaker Recognition Workshop*.
- [52] Yamamoto, S., Nakadai, K., Nakano, M., Tsujino, H., Valin, J.-M., Komatani, K., Ogata, T. et Okuno, H. G. (2006) Real-time robot audition system that recognizes simultaneous speech in the real world. Dans *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*. p. 5333 – 5338.
- [53] Yamamoto, S., Nakadai, K., Valin, J.-M., Rouat, J., Michaud, F., Komatani, K., Ogata, T. et Okuno, H. (2005) Making a robot recognize three simultaneous sentences in real-time. Dans *Proceedings of the International Conference on Intelligent Robots and Systems*. p. 4040 – 4045.
- [54] Zhan, Y., Leung, H., Kwak, K.-C. et Yoon, H. (2009) Automated speaker recognition for home service robots using genetic algorithm and dempster-shafer fusion technique. *IEEE Transactions on Instrumentation and Measurement*, volume 58, n° 9, p. 3058 – 3068.
- [55] Zhang, L., Liu, W. et Langley, R. J. (2010) A class of constrained adaptive beamforming algorithms based on uniform linear arrays. *IEEE Transactions on Signal Processing*, volume 58, n° 7, p. 3916 – 3922.
- [56] Zhang, X., Xiao, X., Wang, H., Suo, H., Zhao, Q. et Yan, Y. (2008) Speaker recognition using a kind of novel phonotactic information. Dans *Chinese Spoken Language Processing*. p. 1 – 4.
- [57] Zhang, Y. et Abdulla, W. (2007) Robust speaker identification in noisy environment using cross diagonal GTF-ICA feature. Dans *Information, Communications Signal Processing*. p. 1 – 4.
- [58] Zotkin, D. et Duraiswami, R. (2004) Accelerated speech source localization via a hierarchical search of steered response power. *IEEE Transactions on Speech and Audio Processing*, volume 12, n° 5, p. 499 – 508.

